FOI
SWEDISH DEFENCE
RESEARCH AGENCY

FOI-R–0119–SE
May 2001
ISSN 1650-1942

Scientific Report

Jonas Persson and Jan Nordström

Discrete Approximations of Electromagnetic Problems

Division of Aeronautics, FFA
SE-172 90 STOCKHOLM

FOI – Swedish Defence Research Agency
Division of Aeronautics, FFA
SE-172 90 STOCKHOLM

FOI-R–0119–SE
May 2001
ISSN 1650-1942

Scientific Report

Jonas Persson and Jan Nordström

Discrete Approximations of Electromagnetic Problems

Division of Aeronautics, FFA
SE-172 90 STOCKHOLM

Abstract

In this report second and higher order methods (the Yee-method, Summation By Parts methods and Finite Element Methods) for transportation of electromagnetic waves are compared. Tests of accuracy, long time-integration and efficiency are performed. We show that the higher order methods in almost every case outperform the second order methods.

Contents

1 Introduction

Maxwells equations describe all electromagnetic phenomena including electromagnetic waves. Such waves appear in many different applications such as microwave-ovens, cellular-phones and radar equipment. Understanding of and ability to solve these equations is crucial when studying these applications. In most cases an analytical solution to these equations cannot be found and numerical methods and computers must be used to find an approximative solution.

In this report we focus on the transportation of electromagnetic waves. Maxwells equations will be solved in time domain and one space dimension. The second order Yee-method will be compared with a second order finite difference method (FDM) using a Summation By Parts (SBP) operator and a second order Finite Element Method (FEM). The second order methods will also be compared with two fourth order methods, one FDM using a fourth order SBP operator and one FEM. A sixth order FDM using a SBP operator will also be considered.

The aim is to gain understanding and a guideline to which method to use in future electromagnetic solvers. Real world problems usually involve the 3D equations, however, the study of the 1D equations provide most of the answers required for choosing a suitable method even for higher dimensional problems. The only additional question is how to scale up the 1D efficiency to 3D. That will be dealt with in section 7.

2 Physics – The governing equations

We are primarily concerned with wave propagation of a transverse electromagnetic wave in a source–free region where both the volume-density of free charges and the density of free currents are equal to zero. We also let the wave propagation take place in a linear, isotropic, homogeneous and nonconducting medium characterized by the permittivity, ϵ and the permeability, μ. This background reduces the Maxwells equations [2] to the following set of equations:

$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t} \quad \nabla \times \mathbf{H} = \epsilon \frac{\partial \mathbf{E}}{\partial t} \tag{1}$$

$$\nabla \cdot \mathbf{E} = 0 \quad \nabla \cdot \mathbf{H} = 0 \tag{2}$$

where $\mathbf{E}$ and $\mathbf{H}$ are vectors of the x, y and $z-$components of the electric and magnetic fields.

2.1 The 1D formulation

If we assume no variations in the electric and magnetic fields in the y– and z–direction, all partial derivatives with respect to y and z will be equal to zero. In 3D the field intensities in Maxwells equations are coupled but when reducing the dimension to 1D they decouple and the resulting set of equations is:

$$\frac{\partial H_y}{\partial t} = \frac{1}{\mu}\left(\frac{\partial E_z}{\partial x}\right) \quad \frac{\partial E_z}{\partial t} = \frac{1}{\epsilon}\left(\frac{\partial H_y}{\partial x}\right), \tag{3}$$

$$\frac{\partial E_y}{\partial t} = \frac{1}{\epsilon}\left(-\frac{\partial H_z}{\partial x}\right) \quad \frac{\partial H_z}{\partial t} = \frac{1}{\mu}\left(-\frac{\partial E_y}{\partial x}\right). \tag{4}$$

The divergence equations (2) are meaningless in one dimension when studying transverse waves. Note that the two first equations, (3) decouple from the last two equations, (4). The equations (3) are usually referred to as the *Transverse Magnetic Mode* (TM) while (4) are referred to as the *Transverse Electric Mode* (TE).

2.2 The model-problem

Since the TM- and TE-modes decouple they can be examined separately. From here on, only the TE-equations will be examined. For simplicity the electric field will be denoted by E and the magnetic field by H (the component indices y and z are dropped).

In this project we consider electromagnetic waves between two *perfect electric conductors* (PEC), e.g. two metal plates with infinite conduction. From the assumption of PEC boundary conditions it follows that the electric field will be zero at the boundaries [2]. This means that we will have electromagnetic standing waves between the two plates.

2.3 Energy estimates of the continuous problem

Let us now write the TE-equations (4) on vector form:

$$\begin{pmatrix} E \\ H \end{pmatrix}_t + \underbrace{\begin{pmatrix} 0 & \frac{1}{\epsilon} \\ \frac{1}{\mu} & 0 \end{pmatrix}}_{=A} \begin{pmatrix} E \\ H \end{pmatrix}_x = 0, \tag{5}$$

where A is the system matrix. The continuous energy rate is obtained by using the energy method [4] on (5) with $(E, H)^T S$ where S is a symmetrization matrix $S = diag(1/\mu, 1/\epsilon)$. The result is:

$$\frac{1}{\mu}\frac{d}{dt}||E||^2 + \frac{1}{\epsilon}\frac{d}{dt}||H||^2 = -\frac{2}{\epsilon\mu}\left(E(1)H(1) - E(0)H(0)\right), \tag{6}$$

which will be of interest later in section 4.

The two equations in (5) can be combined to yield a second order wave-equation

$$U_{tt} = c^2 U_{xx}, \tag{7}$$

where U is either the electric field E or the magnetic field H and the relation $c^2 = 1/\epsilon\mu$ has been used. We get the continuous energy-rate by multiplying (7) with U_t and integrate over $[0, 1]$ yielding

$$\frac{1}{2}\frac{d}{dt}||U_t||^2 + \frac{c^2}{2}\frac{d}{dt}||U_x||^2 = c^2[U_t U_x]_0^1. \tag{8}$$

Note that the energy-rate involves both the time and spatial derivative of U. The estimate (8) will be of interest in section 5.

2.4 Electromagnetic waves

It can be shown (see [2]) that:

The tangential component of an E-field is continuous across an interface between two media.

Since an electric field can not exist inside a perfect electric conductor (i.e. E_{inside} =0) the electric field at the boundary of the perfect electric conductor also has to be zero. In the model problem at hand, an electromagnetic wave impinges at normal incidence at the boundary. Since the electric field is zero at the boundaries, the reflected wave and the impinging wave must have equal magnitude and opposite sign at the boundaries resulting in a standing wave between two perfect electric conductors.[2]

3 The Yee–scheme

In 1966 Kane S. Yee introduced a finite-difference-scheme for solving the Maxwells equations that has been extensively used ever since. It is usually referred to as "The Yee Scheme" or the "FD-TD-scheme" [12]. The scheme proposed by Yee uses a staggered grid in both space and time and is, according to Yee, particularly good for the case with the electrical field equal to zero at the boundary. It uses central leap-frog difference approximations and is second order accurate in both space and time. The resulting time-stepping algorithm is non-dissipative, i.e. the amplitude of the numerical solution will neither grow or decay. [10]

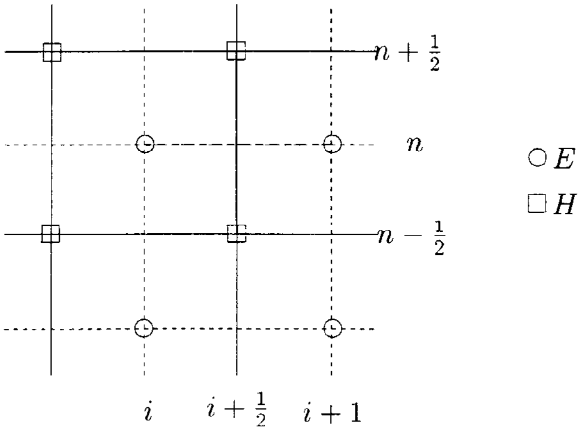
Using a staggered grid means that electric and magnetic field intensities from "half-steps" in the grid are used for the calculations. They are coupled and E-values are used to calculate H-values and vice versa.
The Yee scheme in 1D is:

$$\frac{E_i^{n+1} - E_i^n}{\Delta t} = -\frac{1}{\epsilon}\frac{H_{i+\frac{1}{2}}^{n+\frac{1}{2}} - H_{i-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x} \tag{9}$$

$$\frac{H_{i+\frac{1}{2}}^{n+\frac{1}{2}} - H_{i+\frac{1}{2}}^{n-\frac{1}{2}}}{\Delta t} = -\frac{1}{\mu}\frac{E_{i+1}^n - E_i^n}{\Delta x} \qquad i = 1, \dots, N. \tag{10}$$

GKS-analysis [4] will be performed to analyze the stability of the Yee-scheme to ensure that the boundary condition $E = 0$ does not introduce an instability in the scheme.

Figure 1. The Yee-molecule.



3.1 From the Yee scheme to the discretized wave-equation

To avoid making GKS-analysis on the staggered grid, the Yee scheme is reformulated as two "two-way wave equations", one in E and in one H.

Using equation (9) and (10) we get the normal second order accurate central discretization of the wave equation $u_{tt} = c^2 u_{xx}$. We can solve for either E or H and get:

$$E_i^{n+1} - 2E_i^n + E_i^{n-1} = c^2\lambda^2(E_{i+1}^n - 2E_i^n + E_{i-1}^n), \tag{11}$$

$$H_j^{m+1} - 2H_j^m + H_j^{m-1} = c^2\lambda^2(H_{j+1}^m - 2H_j^m + H_{j-1}^m). \tag{12}$$

Equation (12) has been obtained by a transformation of the indices: $j = i + \frac{1}{2}$, $m = n + \frac{1}{2}$. Note that $\lambda = \frac{\Delta t}{\Delta x}$ and $c = \frac{1}{\epsilon\mu}$ is the speed of light in free space.

3.2 GKS-analysis

The GKS-analysis is a general stability theory for all types of boundary conditions. We start by investigating the difference scheme:

$$E_i^{n+1} - 2E_i^n + E_i^{n-1} = c^2\lambda^2(E_{i+1}^n - 2E_i^n + E_{i-1}^n), \tag{13}$$

with periodic boundary conditions.

We look for solutions to (13) that are “wave-like” in space and make the ansatz:

$$E_i^n = \xi^n e^{-jkx}|_{x=i\Delta x}, \tag{14}$$

where $j = \sqrt{-1}$ is the complex unit and $k = \frac{2\pi}{\lambda}$ is the wavenumber. Using the ansatz (14) in the difference equation (13) yields:

$$\xi^2 - 2A\xi + 1 = 0, \tag{15}$$

where $A = 1 - 2c^2\lambda^2 sin^2(\frac{k\Delta x}{2})$. Equation (15) has a solution of the form $\xi = A \pm \sqrt{A^2 - 1}$. The problem (13) will have a growing unstable solution if $|\xi| > 1$ which can only happen if $|A| > 1$. But $|A| > 1$ require $\frac{c\Delta t}{\Delta x} > 1$. This proves that the difference scheme (13) is stable with periodic boundary conditions under the condition:

$$\Delta t \leq \frac{\Delta x}{c}, \tag{16}$$

and we can continue with the GKS-analysis.

The wave-equation with PEC boundary conditions is:

$$E_i^{n+1} - 2E_i^n + E_i^{n-1} = c^2\lambda^2(E_{i+1}^n - 2E_i^n + E_{i-1}^n) \tag{17}$$

$$E_0^n = 0 \tag{18}$$

$$E_N^n = 0. \tag{19}$$

For the analysis we also require $E \in \mathcal{L}_2(0, 1)$.

We do the ansatz: (i.e. a discrete Laplace transform of the discrete problem)

$$E_i^n = z^n \tilde{E}_i. \tag{20}$$

The Godunov Ryabenkii condition [4] state that a necessary condition for stability is that no solution with $|z| > 1$ exists. With the ansatz (20) in (17) we obtain:

$$(z-1)^2 \tilde{E}_i = c^2 \lambda^2 z(\tilde{E}_{i+1} - 2\tilde{E}_i + \tilde{E}_{i-1}). \tag{21}$$

Equation (21) is the so called **resolvent equation** which have a solution of the form :

$$\tilde{E}_i = \sigma_1^E \varkappa_1^i + \sigma_2^E \varkappa_2^i \tag{22}$$

where σ_1^E and σ_2^E are coefficients to be determined by the boundary conditions. $\varkappa_1$ and $\varkappa_2$ are solutions to the **characteristic equation**:

$$\varkappa(z-1)^2 = c^2 \lambda^2 z(\varkappa - 1)^2. \tag{23}$$

Here $|\varkappa_1| < 1$ and $|\varkappa_2| > 1$. This is realized through the following argument. With $\varkappa = e^{i\omega h}$ we have the same characteristic equation as in the case with periodic boundary conditions and $|\varkappa| = 1$. However, we can't have solutions z with $|z| > 1$ since that scheme is stable. This means that the characteristic equation have no solutions with $|z| > 1 \quad |\varkappa| = 1$. $\varkappa$ is a function of z and the characteristic equation (23) can be written on the form $\varkappa^2 + A\varkappa + 1 = 0$. This means that $\varkappa_1 \varkappa_2 = 1$ and thus we must have $|\varkappa_1| < 1$ and $|\varkappa_2| > 1$.

3.3 Halfspace problems

For the analysis we now divide the problem into two separate parts. One examining $(0 \le x < \infty)$, here called “The right half-space problem” and $(-\infty < x \le 0)$, “The left half-space problem”.

When separating the problem into two parts we must add extra boundary conditions at $\pm\infty$. By demanding

$$\lim_{i \to \infty} |E_i^n| < \infty, \tag{24}$$

$$\lim_{i \to -\infty} |E_i^n| < \infty, \tag{25}$$

we obtain unique solutions to the halfspace problems.

Consider the right half-space problem, $0 \le x < \infty$. Since we can't allow the solution of the resolvent equation (22) to grow as $i \to \infty$, because of the condition (24), we must have $\sigma_2^E = 0$ since $|\varkappa_2| > 1$. The solution to the resolvent equation is therefore $\tilde{E}_i = \sigma_1^E \varkappa_1^i$.

If we now transform the boundary condition (18) we have:

$$E_0^n = z^n \tilde{E}_0 = z^n \sigma_1^E \varkappa_1^0 = 0.$$

The only solution for $|z| > 1 \quad 0 < |\varkappa_1^0| < 1$ is $\sigma_1^E = 0$. Hence $\tilde{E}_i = 0$ and the right half-space problem is stable. Let us now consider the left half-space problem, $-\infty < x \leq 0$. Now σ_2^E must be zero since otherwise $\varkappa_2$ grows as $i \to -\infty$. This means that the solution to the resolvent equation is $\tilde{E}_i = \sigma_2^E \varkappa_2^i$.

The transformation of the boundary condition (19) yields:

$$E_N^n = z^n \tilde{E}_N = z^n \sigma_2^E \varkappa_2^N = 0 \quad \Rightarrow \sigma_2^E = 0,$$

for $|z| > 1$ and $|\varkappa_2| > 1$.

3.4 A borderline case

It is necessary to check the borderline case $|z| = 1$ and $|\varkappa| = 1$ carefully. Solving the characteristic equation (23) for $\varkappa(z)$ yields:

$$\varkappa_{1,2} = 1 + \frac{(z-1)^2}{2c^2\lambda^2 z} \pm \sqrt{\left(1 + \frac{(z-1)^2}{2c^2\lambda^2 z}\right)^2 - 1}. \tag{26}$$

Equation (26) have two possible double roots. The first one is $z = 1$ ($\Rightarrow \varkappa = 1$) which is a double root to the characteristic equation (23). The other one is $z = 1 - 2c^2\lambda^2 \pm i2c\lambda\sqrt{1 - c^2\lambda^2}$ where $|z| = 1$ and $\varkappa = -1 \Rightarrow |\varkappa| = 1$. In any case, the solution is of the form:

$$\tilde{E}_i = (C_1 + iC_2)\varkappa^i. \tag{27}$$

Non-growing solutions require (see (24) and (25)) $C_2 = 0$.

First we consider the right half-space problem, $0 \leq x < \infty$. The transformed boundary-condition is now:

$$E_0^n = z^n \tilde{E}_0 = z^n C_1 \varkappa_1^0 = 0 \quad \Rightarrow C_1 = 0$$

Now consider the left half-space problem, $-\infty < x \leq 0$. In this case:

$$E_N^n = z^n \tilde{E}_N = z^n C_1 \varkappa_2^N = 0$$

which implies that $C_1 = 0$. The above analysis (section 3.3 and 3.4) shows that there are no solutions to the resolvent equation with $|z| \geq 1$, hence the discretization (17)-(19) is stable.

3.5 Stability in the H–field approximation

It was previously shown that it is possible to obtain a discretization with second order central difference approximations of the wave equation, in both the E- and the H–field. In the E–case there is a "natural" boundary condition stating that E is zero on the boundary. When studying the H–field there is no such "natural" boundary condition. One way to introduce the boundary condition $E = 0$ in the H–field is to use the original Yee-molecule (in this case a discretization of $H_t = -\frac{1}{\epsilon}E_x$) with E equal to zero on the boundary. However, this introduces the E–field in the study of the H–field, which complicate the analysis. But there is a way, using the fact that the E–field is zero on the boundary, to eliminate the E–field from the artificial boundary condition in H.

Note that from here on we are using the H-indices, i.e. H-values have integer indices and E-values are at "half"-positions.

3.6 The boundary condition in H

A natural way to impose a boundary condition on the H–field would be to use the Yee-molecules (9) and (10), here in H–coordinates. This an example demonstrating the idea on the left boundary:

$$\frac{H_0^{m+1} - H_0^m}{\Delta t} = -\frac{1}{\epsilon}\frac{E_{1/2}^{m+1/2} - E_{-1/2}^{m+1/2}}{\Delta x}. \tag{28}$$

Take equation (28) in a previous time step and subtract from the equation itself. The result is, using that $E_{-1/2} = 0$ (on the boundary)

$$\frac{H_0^{m+1} - 2H_0^m + H_0^{m-1}}{\Delta t} = -\frac{1}{\epsilon}\frac{E_{1/2}^{m+1/2} - E_{1/2}^{m-1/2}}{\Delta x}. \tag{29}$$

We now use the other Yee-equation

$$\frac{E_{1/2}^{m+1/2} - E_{1/2}^{m-1/2}}{\Delta t} = -\frac{1}{\mu}\frac{H_1^m - H_0^m}{\Delta x}. \tag{30}$$

Solve for $E_{1/2}^{m+1/2} - E_{1/2}^{m-1/2}$ in (30) and use that expression in (29) yielding:

$$\frac{H_0^{m+1} - 2H_0^m + H_0^{m-1}}{\Delta t^2} = \frac{1}{\mu\epsilon}\frac{H_1^m - H_0^m}{\Delta x^2}. \tag{31}$$

With the same kind of derivation it is easy to show that the boundary condition (34) below is equivalent to the Yee-scheme on the right boundary.

3.7 GKS analysis of the H-field approximation

The derived wave-equation including boundary conditions is:

$$\left\{\begin{aligned} &H_j^{m+1} - 2H_j^m + H_j^{m-1} = c^2\lambda^2(H_{j+1}^m - 2H_j^m + H_{j-1}^m) && (32)\\ &H_0^{m+1} - 2H_0^m + H_0^{m-1} = c^2\lambda^2(H_1^m - H_0^m) && (33)\\ &H_N^{m+1} - 2H_N^m + H_N^{m-1} = c^2\lambda^2(H_{N-1}^m - H_N^m) && (34)\end{aligned}\right.$$

$\forall j = 1, \ldots, N-1$. For the analysis we also need that $\lim_{j\to\infty} |H_j^n| < \infty$ and $\lim_{j\to-\infty} |H_j^n| < \infty$. In the same way as for the E-field, the difference equation (32) is stable with periodic boundary conditions under the condition (16).

We make the Laplace ansatz:

$$H_j^m = z^m \tilde{H}_j$$

corresponding to the discrete Laplace-transform and have the **resolvent equation.**

$$(z-1)^2 \tilde{H}_j = c^2 \lambda^2 z (\tilde{H}_{j+1} - 2\tilde{H}_j + \tilde{H}_{j-1}). \tag{35}$$

For $z : |z| > 1$ the resolvent equation have a solution on the form:

$$\tilde{H}_j = \sigma_1^H \varkappa_1^j + \sigma_2^H \varkappa_2^j,$$

where $\varkappa_1$ and $\varkappa_2$ are solutions to the **characteristic equation**:

$$\varkappa(z-1)^2 = c^2 \lambda^2 z (\varkappa - 1)^2. \tag{36}$$

Note that $|\varkappa_1| < 1$ and $|\varkappa_2| > 1$ with the same argument as in the analysis of the E-problem (see section (3.2)).

3.8 Halfspace problems

Since our problem has both a left and a right boundary we divide our analysis into two parts, one concerning $0 \leq x < \infty$ and one concerning $-\infty < x \leq 0$. We must prove stability for both these problems to have stability in the region $0 \leq x \leq 1$.

First consider the right half-space problem. The boundary condition at infinity forces us to let the constant σ_2^H to be zero since $\varkappa_2$ will grow when j tends to infinity. So the solution is : $\tilde{H}_j = \sigma_1^H \varkappa_1^j$. Transforming the boundary condition (33) yields:

$$\sigma_1^H((z-1)^2 - c^2\lambda^2 z(\varkappa_1 - 1)) = 0 \tag{37}$$

which is also referred to as the **determinant condition.**

Assuming $\sigma_1^H \neq 0$ in eq. (37) we can solve for $\varkappa$ and use in eq. (36). The result is: $(z-1)^2 = 0$ which is not fulfilled if $|z| > 1$. This means that the Godunov Ryabenkii condition [4] is fulfilled. For the case $z = 1 \Rightarrow \varkappa = 1$ we perform a perturbation analysis by letting $z = 1 + \delta$ and $\varkappa = 1 + \epsilon$ in eq. (36), where $\delta > 0$ and the sign of ϵ indicates which $\varkappa$ we have. The result is

$$\delta^2 = c^2 \lambda^2 \epsilon^2, \tag{38}$$

which unfortunately gives no information about $\varkappa$.

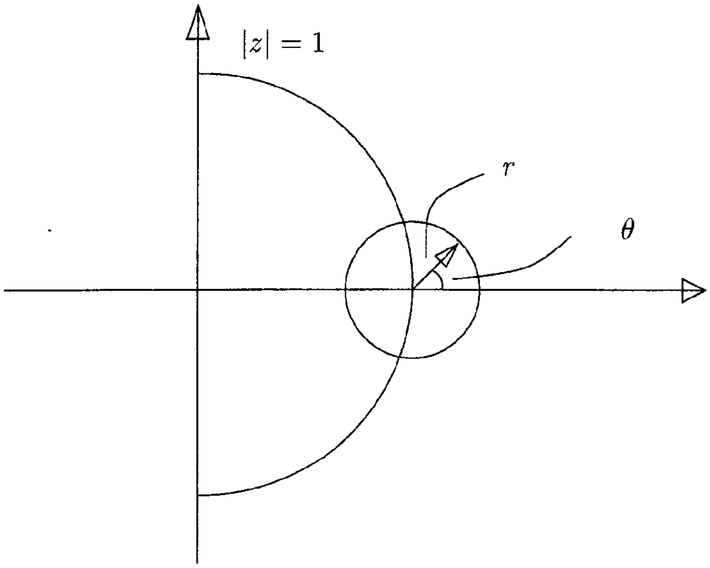


Figure 2. The unit circle with a small circle of radius r at $z = 1$ in the complex plane.

Next we choose $z = 1 + \delta$, $\delta = re^{i\theta}$, where $r \in \mathbb{R}$ and $-\pi/2 \leq \theta \leq \pi/2$ so that we are outside $|z| = 1$ (see fig. 2). The angles $\theta = \pi/2$ and $\theta = -\pi/2$ are good approximations when $r \to 0$. Solving for $\varkappa$ in the transformed boundary condition (37) leads to

$$\varkappa = 1 + \frac{\delta^2}{c^2\lambda^2(1+\delta)} \sim 1 + (\delta/c\lambda)^2 = 1 + \frac{r^2 e^{2i\theta}}{c^2\lambda^2} \qquad (39)$$

where the $\sim$ means a linearization. This means that we have

$$|\varkappa|^2 \approx 1 + \left(\frac{r}{c\lambda}\right)^4 + 2\left(\frac{r}{c\lambda}\right)^2 \cos(2\theta). \qquad (40)$$

When letting $r \to 0$, $|\varkappa|^2 < 1$ if $\cos(2\theta) < 0$. With the previous restriction on θ we now find that $|\varkappa|$ is greater or less than one depending on where in the complex plane we are, see fig. 3. We realize that $|\varkappa| > 1$ if $-\pi/4 \leq \theta \leq \pi/4$ and we have discarded these $\varkappa$ because of the boundary condition at infinity. The dangerous values of z are the ones with $|z| > 1$ and $|\varkappa| < 1$, these could be potential instabilities.

The $\epsilon(\delta)$ that was used in the perturbation analysis can now be identified in (39) as $\varkappa = 1 + (\delta/c\lambda)^2 = 1 + \epsilon_b(\delta)$ where $\epsilon_b(\delta)$ is the function of δ that fulfills the boundary condition. Using this ϵ_b and $z = 1 + \delta$ in the perturbation analysis performed earlier yields

$$\delta^2 = c^2\lambda^2\epsilon_b^2 \quad \Rightarrow \quad 1 = (\delta/c\lambda)^2$$

which is absurd as $\delta \to 0$. This means that there is no solution with $|z| > 1$ and $|\varkappa| < 1$ and the right half-space problem is stable.

For the left half-space problem we consider the characteristic equation (36) and the boundary condition (34). The boundary condition at infinity forces us to let $\sigma_1^H = 0$ since $\varkappa_1$ will grow when j tends to minus infinity.

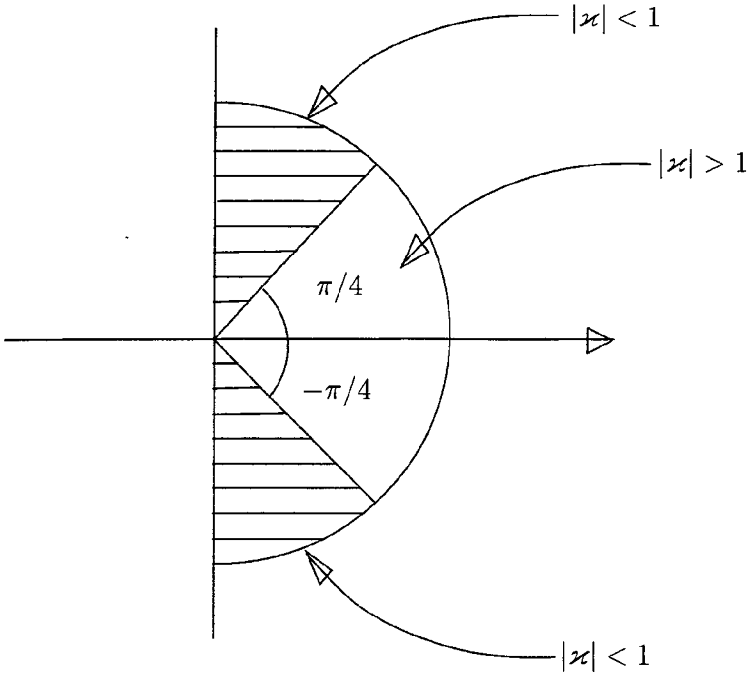


Figure 3. The small circle of radius r in the complex plane . The areas for $|\varkappa| < 1$ and $|\varkappa| > 1$.

This means that the solution is: $\tilde{H}_j = \sigma_2^H \varkappa_2^j$. Transforming the boundary condition (34) yields the **determinant condition**:

$$\sigma_2^H((z-1)^2\varkappa_2 - c^2\lambda^2 z(1-\varkappa_2)) = 0. \tag{41}$$

Dividing equation (41) by $\varkappa_2$ yields the determinant condition $\sigma_2^H((z-1)^2 - c^2\lambda^2 z(\varkappa_2^{-1} - 1)) = 0$ which is the same as for the right half-space problem except that $\varkappa_1$ is replaced by $\varkappa_2^{-1}$. Letting $\tilde{H}_j = \sigma_2^H \varkappa_2^{-j} = \sigma_2^H \left(\varkappa_2^{-1}\right)^j$ and j tends to infinity we realize that the analysis will be the same for the left half-space problem as for the right half-space problem, showing that also the left half-space problem is stable.

We have now shown that the two difference approximations (17)-(19) and (32)-(34) are stable. Since these approximations were derived directly from the Yee-scheme without any approximations or interpolations we conclude that the corresponding problem with the Yee-scheme is also stable.

4 Summation By Parts

4.1 The Semi-Discrete Problem

Using the symmetrization matrix S from section 2.3 and introducing a vector $\omega = (E\ H)^T$ for the unknowns and letting $SA\omega = \hat{A}\omega = F(x)$ transform the continuous system (5) to

$$S\omega_t + F_x = 0. \tag{42}$$

The *Summation By Parts* (SBP) method [7] is a way of constructing an operator for the discretization of the numerical first derivative $\frac{d}{dx}$ in equation (42). It is constructed in such a way that it mimics the integration-by-parts property in the continuous case (see section 2.3). The spatial operator is introduced as:

$$Du = P^{-1}Qu \tag{43}$$

where Du is an approximation of $\frac{d}{dx}E$ or $\frac{d}{dx}H$ and P and Q are matrices. If a spatial operator is of the form (43) and the conditions (i) and (ii) below are full-filled, the operator is referred to as a SBP operator.

(i) The matrix P is symmetric, positive definite and bounded, $\Delta xpI \leq P \leq \Delta xqI$, where $p > 0$ and q are bounded independent of N.

(ii) The matrix Q is almost skew-symmetric with the property: $Q + Q^T = diag(-1, 0, \ldots, 0, 1)$.

We introduce the discrete versions of the vectors of unknowns and matrices in equation (42) by letting

$$\tilde{S} = (I_N \otimes S) \qquad \tilde{F} = (I_N \otimes \hat{A})u \tag{44}$$

$$u = \begin{pmatrix} E_0\ H_0\ E_1\ H_1\ E_2\ H_2 \cdots\ E_N\ H_N \end{pmatrix}^T \tag{45}$$

where $\tilde{S}$ is the discrete version of S and $\tilde{F}$ the discrete version of F. This yields the semi-discrete equation:

$$\tilde{S}u_t + (P^{-1}Q \otimes I_2)\tilde{F} = 0. \tag{46}$$

4.2 An example of a second order accurate SBP-operator

In the case of a second order operator, the matrices P and Q are:

$$P = \Delta x \begin{pmatrix} \frac{1}{2} & & & & & \\ & 1 & & 0 & & \\ & & 1 & & & \\ & & & \ddots & & \\ & 0 & & & 1 & \\ & & & & & \frac{1}{2} \end{pmatrix}, \quad Q = \begin{pmatrix} -\frac{1}{2} & \frac{1}{2} & & & & \\ -\frac{1}{2} & 0 & \frac{1}{2} & & 0 & \\ 0 & -\frac{1}{2} & 0 & \frac{1}{2} & & \\ & & & \ddots & & \\ & 0 & & -\frac{1}{2} & 0 & \frac{1}{2} \\ & & & & -\frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

The resulting discrete operator $(P^{-1}Q \otimes I_2)$ is:

$$(P^{-1}Q \otimes I_2) = \frac{1}{2\Delta x}\begin{pmatrix} -2 & 0 & 2 & 0 & & & & & \\ 0 & -2 & 0 & 2 & 0 & & & & \\ -1 & 0 & 0 & 0 & 1 & 0 & & & \\ 0 & -1 & 0 & 0 & 0 & 1 & 0 & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & 0 & -1 & 0 & 0 & 0 & 1 & 0 \\ & & & 0 & -1 & 0 & 0 & 0 & 1 \\ & & & & 0 & -2 & 0 & 2 & 0 \\ & & & & & & -2 & 0 & 2 \end{pmatrix}.$$

The fourth and sixth order accurate operators are more complex and can be further studied in [7].

4.3 Strict stability

By multiplying equation (46) by $u^T(P \otimes I_2)$ from the left hand side and adding the transpose of the resulting equation, we get

$$u^T(P \otimes I_2)\tilde{S}u_t + u^T(Q \otimes I_2)\tilde{F} + u_t^T\tilde{S}(P \otimes I_2)u + \tilde{F}^T(Q^T \otimes I_2)u = 0.$$

The two terms with time derivatives form the time derivative of a norm, i.e.

$$u^T(P \otimes I_2)\tilde{S}u_t + u_t^T\tilde{S}(P \otimes I_2)u = \frac{d}{dt}\left(u^T\tilde{S}(P \otimes I_2)u\right) = \frac{d}{dt}||u||^2_{(P\otimes S)}.$$

Note that $\tilde{S}(P \otimes I_2) = (I_N \otimes S)(P \otimes I_2) = (P \otimes S)$ and that $(P \otimes I_2)(I_N \otimes S) = (I_N \otimes S)(P \otimes I_2)$. The matrix S is diagonal with elements strictly positive and P is positive definite so $(P \otimes S)$ truly defines a norm.

Using our definition of $\tilde{F}$ in (44) and letting $B = Q + Q^T$ we have:

$$\frac{d}{dt}||u||^2_{(P\otimes S)} = -u^T(B \otimes \hat{A})u = -\frac{2}{\epsilon\mu}(E_N H_N - E_0 H_0)$$

which corresponds to the continuous energy rate in equation (6).

4.4 SAT boundary conditions

Instead of imposing the boundary conditions directly (which might destroy the SBP property) the *SAT* (Simultaneous Approximation Term) method will be used [1]. An extra term, proportional to the difference between the discrete value and the boundary term, is added to the operator and the differential equation can then be solved in all points, also at the boundaries. The extra term does not lower the overall accuracy since it vanishes upon substitution of the exact value.

Let the boundary conditions be arbitrary in the first part of the derivation:

$$u_0 = g_0 = (g_{0E}\ g_{0H})^T \text{ and } u_N = g_N = (g_{NE}\ g_{NH})^T$$

Note that one can only impose one boundary condition at each boundary, e.g. g_{0E} or g_{0H} at the left boundary. The resulting SBP-formulation including the SAT term for boundary conditions is

$$\begin{aligned}\tilde{S}u_t + (P^{-1}Q \otimes I_2)\tilde{F} = (P^{-1} \otimes \hat{A})M_0(u - (W_0 \otimes g_0)) +& \\ (P^{-1} \otimes \hat{A})M_N(u - (W_N \otimes g_N)),&\end{aligned} \tag{47}$$

where

$$M_0 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & 0 \end{pmatrix}_{(N+1)\times(N+1)} \otimes \begin{pmatrix} \sigma_{L_1} & 0 \\ 0 & \sigma_{L_2} \end{pmatrix}_{(2\times 2)} \tag{48}$$

$$M_N = \begin{pmatrix} 0 & & \cdots & 0 \\ \vdots & \ddots & \vdots & 0 \\ 0 & & 0 & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}_{(N+1)\times(N+1)} \otimes \begin{pmatrix} \sigma_{R_1} & 0 \\ 0 & \sigma_{R_2} \end{pmatrix}_{(2\times 2)} \tag{49}$$

$$W_0 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{(N+1)}, \quad W_N = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}_{(N+1)}. \tag{50}$$

The vectors are defined as:

$$u = (u_0,\, u_1,\, u_2, \cdots, u_N)^T \quad \text{and} \quad u_i = (E_i,\, H_i).$$

The energy method applied to equation (47) leads to,

$$\begin{aligned}\frac{d}{dt}\|u\|^2_{(P\otimes S)} = -u^T(B \otimes \hat{A})u + u^T(I_N \otimes \hat{A})M_0(u - (W_0 \otimes g_0)) +& \\ +u^T(I_N \otimes \hat{A})M_N(u - (W_N \otimes g_N)) +& \\ +(u - (W_0 \otimes g_0))^T M_0^T (I_N \otimes \hat{A})u +& \\ +(u - (W_N \otimes g_N))^T M_N^T (I_N \otimes \hat{A})u.&\end{aligned} \tag{51}$$

The sparse structure of the matrices results in only a few non–zero elements from the SAT-terms in equation (47), actually there are just two non-zero

elements in each term.
Letting

$$\tilde{M}_0 = \begin{pmatrix} 0 & \frac{\sigma_{L_2}}{\epsilon\mu} \\ \frac{\sigma_{L_1}}{\epsilon\mu} & 0 \end{pmatrix} \text{ be the upper left } (2 \times 2)\text{-matrix in } (I_N \otimes \hat{A})M_0$$

$$\tilde{M}_N = \begin{pmatrix} 0 & \frac{\sigma_{R_2}}{\epsilon\mu} \\ \frac{\sigma_{R_1}}{\epsilon\mu} & 0 \end{pmatrix} \text{ be the lower right } (2 \times 2)\text{-matrix in } (I_N \otimes \hat{A})M_N$$

and using that $B = Q + Q^T$ we have that equation (51) becomes

$$\begin{aligned} \frac{d}{dt}\|u\|^2_{(P\otimes S)} =& u_0^T(\hat{A} + \tilde{M}_0 + \tilde{M}_0^T)u_0 - u_N^T(\hat{A} - \tilde{M}_N - \tilde{M}_N^T)u_N - \\ &- u_0^T\tilde{M}_0 g_0 - g_0\tilde{M}_0^T u_0 - u_N^T\tilde{M}_N g_N - g_N\tilde{M}_N^T u_N. \end{aligned} \tag{52}$$

Since we can not impose more than one boundary condition at each boundary we must now let one of σ_{L_1} and σ_{L_2} at the left boundary and one of σ_{R_1} and σ_{R_2} at the right boundary be zero. This means imposing a boundary condition on either E or H at each boundary. The first two terms in (52) are indefinite and can be eliminated if

$$\tilde{A} + \tilde{M}_0 + \tilde{M}_0^T = \frac{1}{\epsilon\mu}\begin{pmatrix} 0 & 1 + \sigma_{L_1} + \sigma_{L_2} \\ 1 + \sigma_{L_1} + \sigma_{L_2} & 0 \end{pmatrix} = 0 \tag{53}$$

$$\tilde{A} - \tilde{M}_N - \tilde{M}_N^T = \frac{1}{\epsilon\mu}\begin{pmatrix} 0 & 1 - \sigma_{R_1} - \sigma_{R_2} \\ 1 - \sigma_{R_1} - \sigma_{R_2} & 0 \end{pmatrix} = 0. \tag{54}$$

In the model-problem we only have boundary conditions on the electric field E and thus $\sigma_{L_2} = \sigma_{R_2} = 0$ is an appropriate choice of two of the penalty parameters. Equation (53) and (54) then force us to let $\sigma_{L_1} = -1$ and $\sigma_{R_1} = 1$. The remaining terms in (52) become,

$$\begin{aligned} -u_0^T\tilde{M}_0 g_0 &= -\begin{pmatrix} E_0 & H_0 \end{pmatrix}\begin{pmatrix} 0 & 0 \\ -\frac{1}{\epsilon\mu} & 0 \end{pmatrix}\begin{pmatrix} g_{0E} \\ g_{0H} \end{pmatrix} = \frac{1}{\epsilon\mu}H_0 g_{0E} = 0, \\ -g_0\tilde{M}_0^T u_0 &= -\begin{pmatrix} g_{0E} & g_{0H} \end{pmatrix}\begin{pmatrix} 0 & -\frac{1}{\epsilon\mu} \\ 0 & 0 \end{pmatrix}\begin{pmatrix} E_0 \\ H_0 \end{pmatrix} = \frac{1}{\epsilon\mu}H_0 g_{0E} = 0, \\ -u_N^T\tilde{M}_N g_N &= -\begin{pmatrix} E_N & H_N \end{pmatrix}\begin{pmatrix} 0 & 0 \\ \frac{1}{\epsilon\mu} & 0 \end{pmatrix}\begin{pmatrix} g_{NE} \\ g_{NH} \end{pmatrix} = -\frac{1}{\epsilon\mu}H_N g_{NE} = 0, \\ -g_N\tilde{M}_N^T u_N &= -\begin{pmatrix} g_{NE} & g_{NH} \end{pmatrix}\begin{pmatrix} 0 & \frac{1}{\epsilon\mu} \\ 0 & 0 \end{pmatrix}\begin{pmatrix} E_N \\ H_N \end{pmatrix} = -\frac{1}{\epsilon\mu}H_N g_{NE} = 0, \end{aligned}$$

since $E = 0$ at the boundary, i.e. $g_{0E} = g_{NE} = 0$. The energy rate is thus identical to zero and there is no dissipation. Note that the boundary terms g_{0H} and g_{NH} do not appear in the remaining terms.

4.5 Time integration method for the SBP

For the second, fourth and the sixth order accurate SBP-operator in space, the classical fourth order Runge-Kutta method was used for the time integration (see any textbook in Numerical Analysis).

5 Finite Element Methods

Maxwells equations can be combined to yield the two–way wave equation $u_{tt} = c^2 u_{xx}$, where u is either the electric field E or the magnetic field H. This is easily done by first taking a spatial derivative of the first equation and a time derivative of the second equation in (4).

Now we want to formulate a FEM approximation for the two-way wave-equation:

$$\begin{cases} u_{tt} = c^2 u_{xx} & x \in [0,1] \quad t \geq 0 \\ u(0,t) = u(1,t) = 0 & \\ u(x,0) = \sin(2\pi n x) \ , \ u_t(x,0) = 0 & \end{cases} \tag{55}$$

where n is the number of wavelengths in the domain and the intialdata are derived from the exact solution $u(x,t) = \sin(2\pi n x)\sin(2\pi n t + \pi/2)$.

5.1 Continuous case – a variational formulation

To construct a continuous variational formulation of the wave equation in (55) let V be a linear space:

$V = \{v : v$ is a continuous function on $[0,1]$,
v' is piecewise continuous and limited function on $[0,1]$.
And $v(0) = v(1) = 0\}$

Now multiply the wave equation in (55) with a test function $v \in V$ and integrate over $[0,1]$.

$$\int_0^1 v u_{tt} dx = \int_0^1 v c^2 u_{xx} dx = -c^2 \int_0^1 v_x u_x dx \tag{56}$$

We introduce the notation

$$\int_0^1 v w dx = (v,w)$$

so that equation (56) becomes

$$(v, u_{tt}) + c^2 (v_x, u_x) = 0\,. \tag{57}$$

The continuous **Variational formulation** can be stated as:

$$\begin{cases} \text{Find } u \in V \text{ such that:} \\ \\ (v, u_{tt}) + c^2(v_x, u_x) = 0 \ \forall v \in V \end{cases} \tag{58}$$

5.2 A semi-discrete approximation

Perform a semi-discretization in space by introducing a linear subspace V_h. To create the linear subspace $V_h \subset V$ (e.g. using piecewise linear polynomials) we introduce basis-functions $\varphi_i(x) \in V$. The subspace V_h is spanned by $\{\varphi_i\}_{i=1,\dots,N}$, i.e. V_h is a linear subspace of V of dimension N.

A function $v \in V_h$ can be expressed as a linear combination of the basis functions:

$$v(x) = \sum_{i=1}^{N} \xi_i \varphi_i(x) \quad x \in [0,1].$$

A finite dimensional variational formulation is

$$\begin{cases} \text{Find } u_h \in V_h \text{ such that:} \\ (v, (u_h)_{tt}) + c^2(v_x, (u_h)_x) = 0 \ \forall v \in V_h. \end{cases} \tag{59}$$

Equation (59) is true $\forall v \in V_h$ and thus also true for all $\varphi_i \in V_h$, i.e.

$$(\varphi_i, (u_h)_{tt}) + c^2((\varphi_i)_x, (u_h)_x) = 0 \quad \forall i = 1, \dots, N. \tag{60}$$

V_h is spanned by the basis function φ_i and if $u_h \in V_h$, u_h can be obtained as

$$u_h(x,t) = \sum_{j=1}^{N} \xi_j \varphi_j(x), \qquad \xi_j(t) = u_h(x_j, t), \tag{61}$$

and we have

$$\sum_{j=1}^{N} (\xi_j)_{tt}(\varphi_i, \varphi_j) + c^2 \sum_{j=1}^{N} \xi_j((\varphi_i)_x, (\varphi_j)_x) = 0 \quad \forall i = 1, \dots, N \tag{62}$$

since the basis-functions are time independent. Equation (62) defines a system of equations and can be written as

$$M\xi_{tt} + c^2 K\xi = 0, \tag{63}$$

where $M_{ij} = (\varphi_i, \varphi_j)$ and $K_{ij} = ((\varphi_i)_x, (\varphi_j)_x)$.

5.3 A second order approximation

When constructing basis–functions for a second-order approximation on an equidistant grid, first order polynomials suffice. Let $x_i : \quad i = 0, \dots, N+1$, define the nodes and let h_i define the spaces between the nodes, $h_i = x_i - x_{i-1} : \quad i = 1, \dots, N+1$.

The basis-functions are, see fig. 4,

$$\varphi_i(x) = \begin{cases} \frac{x - x_{i-1}}{h_i} \ x_{i-1} \le x \le x_i \\ \qquad\qquad\qquad\qquad i = 1, \dots. N. \\ \frac{x_{i+1} - x}{h_{i+1}} \ x_i < x \le x_{i+1} \end{cases} \tag{64}$$

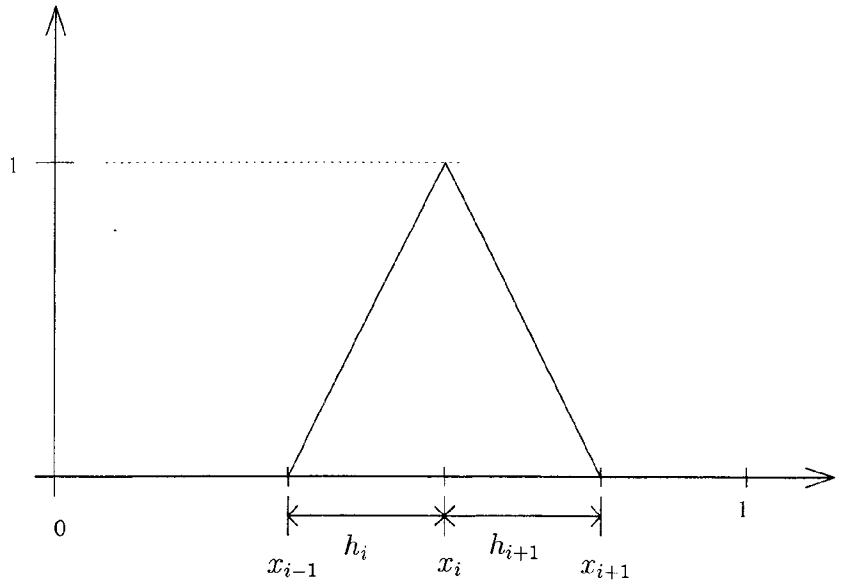


Figure 4. Basisfunctions for a second order approximation.

The basis–function φ_i is 1 in x_i and 0 in the other grid-points. To fulfill the boundary–conditions we omit the left- and rightmost basis–functions φ_0 and φ_{N+1}. The basis–function φ_0 and φ_{N+1} will then naturally be zero on the boundary and can therefore be excluded from the calculations. The derivative of the basis–functions are:

$$\varphi_i'(x) = \begin{cases} h_i^{-1} & x_{i-1} \leq x \leq x_i \\ -h_{i+1}^{-1} & x_i < x \leq x_{i+1} \end{cases} \quad i = 1, \ldots, N\,. \tag{65}$$

Because of local support of the basis–functions φ_i and φ_i' the structure of

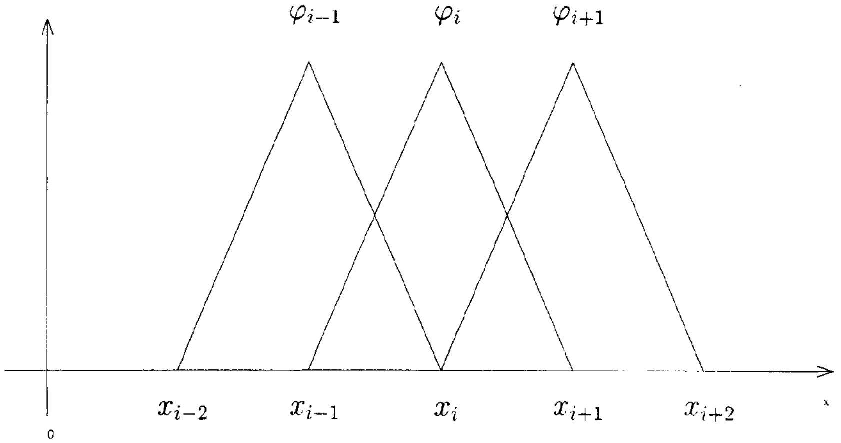


Figure 5. Basis-functions and overlap

the Matrices M and K will be tridiagonal with the non–zero elements:

$$K_{i,i} = \int\limits_0^1 \varphi_i'\varphi_i' dx = \int\limits_{x_{i-1}}^{x_{i+1}} \varphi_i'\varphi_i' dx = h_i^{-1} + h_{i+1}^{-1} \tag{66}$$

$$K_{i,i-1} = K_{i-1,i} = \int\limits_{x_{i-1}}^{x_i} \varphi_{i-1}'\varphi_i' dx = \int\limits_{x_{i-1}}^{x_i} -h_i^{-1}h_i^{-1} dx = -h_i^{-1}. \tag{67}$$

For an equidistant grid, the matrix K is:

$$K = h^{-1}\begin{pmatrix} 2 & -1 & 0 & & \dots & 0 \\ -1 & 2 & -1 & 0 & & \vdots \\ 0 & -1 & 2 & -1 & & \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ & & & -1 & 2 & -1 \\ 0 & \dots & & 0 & -1 & 2 \end{pmatrix}.$$

The nonzero elements in the matrix M are:

$$M_{i,i} = \int\limits_0^1 \varphi_i\varphi_i dx = \int\limits_{x_{i-1}}^{x_{i+1}} \varphi_i\varphi_i dx = \frac{h_i}{3} + \frac{h_{i+1}}{3} \tag{68}$$

$$M_{i,i-1} = M_{i-1,i} = \int\limits_{x_{i-1}}^{x_i} \varphi_{i-1}\varphi_i dx = \int\limits_{x_{i-1}}^{x_i} \frac{(x_i - x)(x - x_{i-1})}{h_i^2} dx = \frac{h_i}{6}. \tag{69}$$

The matrix M for an equidistant grid is:

$$M = \frac{h}{3}\begin{pmatrix} 2 & \frac{1}{2} & 0 & & \dots & 0 \\ \frac{1}{2} & 2 & \frac{1}{2} & 0 & & \vdots \\ 0 & \frac{1}{2} & 2 & \frac{1}{2} & & \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ & & & \frac{1}{2} & 2 & \frac{1}{2} \\ 0 & \dots & & 0 & \frac{1}{2} & 2 \end{pmatrix}.$$

Note that both K and M represent the spatial operator in the inner part of $[0, 1]$.

In this section we have so far discretized space and kept time continuous. The result is a system of ordinary differential equations.

$$M\xi_{tt} + c^2 K\xi = 0. \tag{70}$$

By multiplying equation (70) by ξ_t^T and using the structure of M and K and (61) we get:

$$\frac{1}{2}\frac{d}{dt}||(u_h)_t||^2 + \frac{c^2}{2}\frac{d}{dt}||(u_h)_x||^2 = 0$$

which corresponds to the continuous energy rate in (8).

A second order time-integrator is needed for the time-integration of the second order FEM-approximation. From here on we denote ξ by u and choose the "standard" second-order central difference approximation to integrate u_{tt} in time, i.e. $(u^{n+1} - 2u^n + u^{n-1})/\Delta t^2 \approx u_{tt}$. A fully discrete version of equation (55) with $c = 1$ (a scaling of ϵ and μ) is then:

$$M\frac{u^{n+1} - 2u^n + u^{n-1}}{\Delta t^2} + Ku^n = 0. \tag{71}$$

Since the system defined by (71) is symmetric and tri-diagonal there are fast solvers available. Gaussian elimination can be performed without row or column pivoting since the matrix M is diagonally dominant. First the system is converted to upper-triangular form and then one forward- and one backward-substitution is needed. For details see e.g. [5].

To determine the stability region, we write (71) on a one–step form, i.e.

$$\underbrace{\begin{pmatrix} M & 0 \\ 0 & I \end{pmatrix}}_{=\tilde{M}} \begin{pmatrix} u^{n+1} \\ u^n \end{pmatrix} = \underbrace{\begin{pmatrix} 2I - \Delta t^2 K & -M \\ I & 0 \end{pmatrix}}_{=\tilde{K}} \begin{pmatrix} u^n \\ u^{n-1} \end{pmatrix} \tag{72}$$

$$\underbrace{\begin{pmatrix} u^{n+1} \\ u^n \end{pmatrix}}_{=U^{n+1}} = \underbrace{\tilde{M}^{-1}\tilde{K}}_{=A} \begin{pmatrix} u^n \\ u^{n-1} \end{pmatrix}. \tag{73}$$

Suppose A is diagonalizable with $A = X\Lambda X^{-1}$ for a diagonal matrix Λ and a set of eigenvectors X, equation (73) then becomes

$$U^{n+1} = X\Lambda X^{-1}U^n = X\Lambda^n X^{-1}U^0 \tag{74}$$

and we see that the eigenvalues to A must be distinct and ≤ 1 for stability. To find the eigenvalues to A, let an eigenvector be $(a\,,b)^T$ with a and b being vectors of length N that correspond to the size of M and K. The eigenvalue-problem is then

$$A\begin{pmatrix} a \\ b \end{pmatrix} = \lambda\begin{pmatrix} a \\ b \end{pmatrix} \Leftrightarrow \begin{cases} (2I - \Delta t^2 M^{-1}K)a - b = \lambda a \\ Ia = \lambda b. \end{cases} \tag{75}$$

Using equation (75) we find that

$$-\Delta t^2 M^{-1}Ka = \underbrace{(\lambda + \frac{1}{\lambda} - 2)}_{=-\mu}\, a \tag{76}$$

where $-\mu$ is an eigenvalue to $-\Delta t^2 M^{-1}K$. The eigenvalue-problem (75) for the matrix in (73) has been transformed to the smaller eigenvalue problem (76) for the matrix $M^{-1}K$. From (76) we have

$$\lambda_{1,2} = 1 - \frac{\mu}{2} \pm \sqrt{\left(1 - \frac{\mu}{2}\right)^2 - 1}. \tag{77}$$

The eigenvalues λ can be either real or complex depending on the sign of $\left(1-\frac{\mu}{2}\right)^2-1=C$. We observe that $C<0$ in $0<\mu<4$, which implies $|\lambda|=1$ which is permitted for stability reasons. Let us now study the absolute value of λ_1 and λ_2 in μ=[-0.5, 4.5]. Fig. 6 show that outside $0<\mu<4$ we have $|\lambda|>1$ for one of the two roots. We also note that $\mu=0 \Rightarrow \lambda=1$ and $\mu=4 \Rightarrow \lambda=-1$. $\mu\geq 0$ must hold since $\Delta t^2x^TKx=\mu x^TMx$, $x^TKx\geq 0$ and $x^TMx\geq 0$ for an eigenvector x. Conclusion: $0\leq\mu\leq 4=\mu_0$ is required for $|\lambda|\leq 1$ and stability.

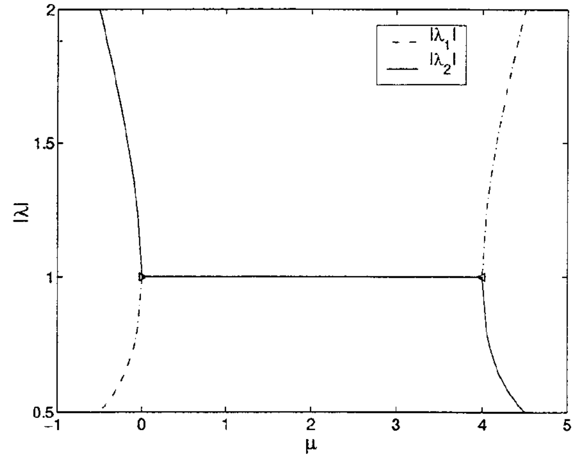


Figure 6. The absolute value of λ_1 and λ_2.

To find an estimate of the stability limit we must estimate the largest eigenvalue μ of $\Delta t^2M^{-1}K$. To get an idea we assume that the problem is periodic (a major simplification) and study how the matrices M and K act on a vector u. We do the ansatz: $u_j=e^{ikx_j}\hat{u}(t)$. For any row j in the inner part of M and K, we have

$$\sum_{l=1}^{N}M_{jl}u_l=\frac{\Delta x}{3}(u_{j-1}/2+2u_j+u_{j+1}/2)=$$
$$=u_j\frac{\Delta x}{3}(2+\cos(k\Delta x))\quad j=1,\dots,N$$
$$\sum_{l=1}^{N}K_{jl}u_l=\frac{1}{\Delta x}(-u_{j-1}+2u_j-u_{j+1})=$$
$$=\frac{2}{\Delta x}u_j(1-\cos(k\Delta x))\quad j=1,\dots,N.$$

Note that the two matrices have the same eigenvectors u . This means that they can be diagonalized simultaneously in the same basis of eigenvectors.

If we let X be a set of such eigenvectors we have

$$MX=\Lambda_MX \Rightarrow M=X\Lambda_MX^{-1}$$
$$KX=\Lambda_KX \Rightarrow K=X\Lambda_KX^{-1},$$

and thus $M^{-1}K=X\Lambda_M^{-1}\Lambda_KX^{-1}$ where Λ_M and Λ_K are diagonal matrices with the eigenvalues of M and K respectively. Note that $\Lambda_M=\frac{\Delta x}{3}(2+$

$\cos(k\Delta x))$ and $\Lambda_K = \frac{2}{\Delta x}(1 - \cos(k\Delta x))$ and the eigenvalues μ are then

$$\mu = \frac{\frac{2}{\Delta x}(1 - \cos(k\Delta x))}{\frac{\Delta x}{3}(2 + \cos(k\Delta x))} = \frac{6}{\Delta x^2} \frac{(1 - \cos(k\Delta x))}{(2 + \cos(k\Delta x))}. \tag{78}$$

Maximum of (78), μ_{max}, is $\frac{12}{\Delta x^2}$ when $\cos(k\Delta x) = -1$. For stability we must have $\Delta t^2 \mu_{max} \le \mu_0$, i.e.

$$\frac{\Delta t^2}{\Delta x^2} 12 \le 4 \quad \Rightarrow \quad \frac{\Delta t}{\Delta x} \le \frac{1}{\sqrt{3}}.$$

The stability criterion for the second order FEM approximation with periodic boundary conditions give an estimate of the timestep one can use in the non-periodic case.

5.4 A fourth order approximation

To achieve a fourth order approximation in space, Lagrange interpolation polynomials of order three has been used as basis-functions, see [5] .

The Lagrange interpolation polynomials for an element e can generally be written as:

$$\begin{aligned}
\varphi^{e_1} &= \frac{(x - x_2^e)(x - x_3^e)(x - x_4^e)}{(x_1^e - x_2^e)(x_1^e - x_3^e)(x_1^e - x_4^e)} \\
\varphi^{e_2} &= \frac{(x - x_1^e)(x - x_3^e)(x - x_4^e)}{(x_2^e - x_1^e)(x_2^e - x_3^e)(x_2^e - x_4^e)} \\
\varphi^{e_3} &= \frac{(x - x_1^e)(x - x_2^e)(x - x_4^e)}{(x_3^e - x_1^e)(x_3^e - x_2^e)(x_3^e - x_4^e)} \\
\varphi^{e_4} &= \frac{(x - x_1^e)(x - x_2^e)(x - x_3^e)}{(x_4^e - x_1^e)(x_4^e - x_2^e)(x_4^e - x_3^e)}
\end{aligned}$$

We let the finite element e be specified by four nodes at equidistant distance. On every element there are four basis-functions each having the value 1 in one node and the value zero in the other nodes. The Lagrange interpolation polynomials have exactly this property. Note that a third order polynomial is uniquely determined by the specification of four points.

In figure 7, the four basis-functions are displayed. The nodes are located in [-1 -1/3 1/3 1] and we see e.g. that basis-function 1 have the value one in -1 and is zero in -1/3, 1/3 and 1.

The next step is to compute the matrix elements in the matrix M. Since the basis-functions are local to each element we first compute a local matrix M_l and then use the *assembly technique* (see e.g. Claes Johnson [6]) to form the matrix M. We compute, for each basis-function, the contribution from the basis-function overlapping itself and the contribution from the basis-function overlapping the other basis-functions on the element. The matrix K is computed in the same way as M by first computing a local matrix K_l where the matrix elements are the contributions from the derivatives

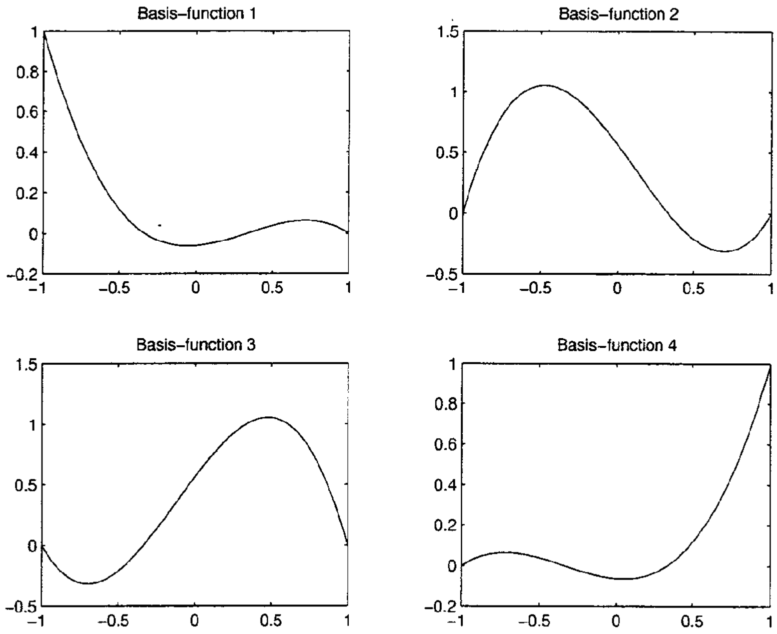


Figure 7. Basis-functions of order 3 scaled to range from -1 to 1

of the basis-functions overlapping themselves and the other basis-functions of the element.

When an equidistant grid is used all local matrices will be the same. The products of the Lagrange interpolation polynomials (and the product of the derivatives of the polynomials) can be integrated exactly. This was done in [5] and the result is used in this work. The local M_l and K_l-matrices for element e are:

$$M_l = \begin{pmatrix} \int_0^1 \varphi^{e_1}\varphi^{e_1}dx & \int_0^1 \varphi^{e_1}\varphi^{e_2} & \int_0^1 \varphi^{e_1}\varphi^{e_3} & \int_0^1 \varphi^{e_1}\varphi^{e_4} \\ \int_0^1 \varphi^{e_2}\varphi^{e_1} & \int_0^1 \varphi^{e_2}\varphi^{e_2} & \int_0^1 \varphi^{e_2}\varphi^{e_3} & \int_0^1 \varphi^{e_2}\varphi^{e_4} \\ \int_0^1 \varphi^{e_3}\varphi^{e_1} & \int_0^1 \varphi^{e_3}\varphi^{e_2} & \int_0^1 \varphi^{e_3}\varphi^{e_3} & \int_0^1 \varphi^{e_3}\varphi^{e_4} \\ \int_0^1 \varphi^{e_4}\varphi^{e_1} & \int_0^1 \varphi^{e_4}\varphi^{e_2} & \int_0^1 \varphi^{e_4}\varphi^{e_3} & \int_0^1 \varphi^{e_4}\varphi^{e_4} \end{pmatrix}$$

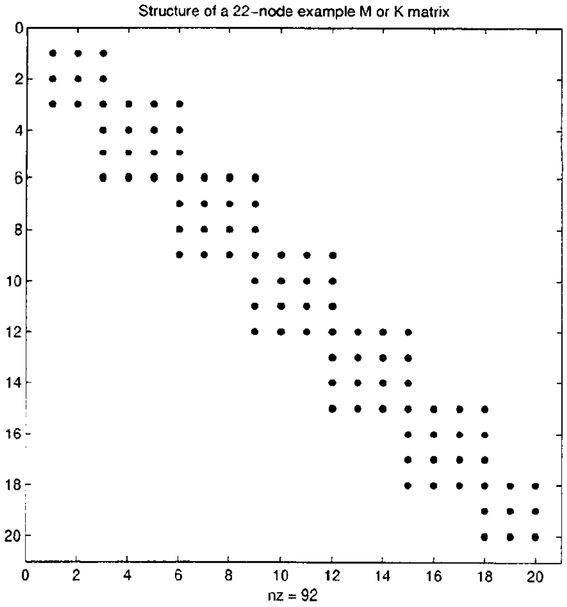
$$= 3\Delta x \begin{pmatrix} \frac{8}{105} & \frac{33}{560} & \frac{-3}{140} & \frac{19}{1680} \\ \frac{33}{560} & \frac{27}{70} & \frac{-27}{560} & \frac{-3}{140} \\ \frac{-3}{140} & \frac{-27}{560} & \frac{27}{70} & \frac{33}{560} \\ \frac{19}{1680} & \frac{-3}{140} & \frac{33}{560} & \frac{8}{105} \end{pmatrix};$$

$$K_l = \begin{pmatrix} \int_0^1 \varphi_x^{e_1}\varphi_x^{e_1} & \int_0^1 \varphi_x^{e_1}\varphi_x^{e_2} & \int_0^1 \varphi_x^{e_1}\varphi_x^{e_3} & \int_0^1 \varphi_x^{e_1}\varphi_x^{e_4} \\ \int_0^1 \varphi_x^{e_2}\varphi_x^{e_1} & \int_0^1 \varphi_x^{e_2}\varphi_x^{e_2} & \int_0^1 \varphi_x^{e_2}\varphi_x^{e_3} & \int_0^1 \varphi_x^{e_2}\varphi_x^{e_4} \\ \int_0^1 \varphi_x^{e_3}\varphi_x^{e_1} & \int_0^1 \varphi_x^{e_3}\varphi_x^{e_2} & \int_0^1 \varphi_x^{e_3}\varphi_x^{e_3} & \int_0^1 \varphi_x^{e_3}\varphi_x^{e_4} \\ \int_0^1 \varphi_x^{e_4}\varphi_x^{e_1} & \int_0^1 \varphi_x^{e_4}\varphi_x^{e_2} & \int_0^1 \varphi_x^{e_4}\varphi_x^{e_3} & \int_0^1 \varphi_x^{e_4}\varphi_x^{e_4} \end{pmatrix} =$$

$$= \frac{1}{3\Delta x}\begin{pmatrix} \frac{37}{10} & \frac{-189}{40} & \frac{27}{20} & \frac{-13}{40} \\ \frac{-189}{40} & \frac{54}{5} & \frac{-297}{40} & \frac{27}{20} \\ \frac{27}{20} & \frac{-297}{40} & \frac{54}{5} & \frac{-189}{40} \\ \frac{-13}{40} & \frac{27}{20} & \frac{-189}{40} & \frac{37}{10} \end{pmatrix}.$$

The four times four matrices are then assembled to form the matrix M and the matrix K by adding the local matrices into the global ones. The local support of the basis-functions yield a seven-diagonal structure of M and K (see figure (8)). Note that the basis-functions φ^{e_1} and φ^{e_4} have the value one at the boundary of the element and will be connected to the surrounding basis-functions and that M and K are symmetric matrices.

Figure 8. Structure of M and K for a 22 node example.



The result is the system of ordinary differential equations:

$$Mu_{tt} + Ku = 0, \tag{79}$$

where u is a vector of the unknown E–values. Note that the boundary condition, $E = 0$, is imposed by removal of the outermost basis-functions (we have only 3×3 blocks in the left upper and right lower corners of M and K). This is sufficient since the other basis-functions in the outermost finite elements are zero in these gridpoints.

A fourth order accurate time integrator is needed to match the fourth order accurate finite element spatial approximation. We have chosen to use the Numerov method as described in [9]:

$$\hat{M} y_{tt} = F(y) \tag{80}$$
$$\hat{M}\frac{y^{n+1}-2y^{n}+y^{n-1}}{\Delta t^2} = \frac{1}{12}\left(F(y^{n+1})+F(y^{n-1})\right)+\frac{5}{6}F(y^{n}).$$

The Numerov method is implicit but it adds no extra complexity to the scheme since the finite element approximation in space also requires the solution of a system of equations. The numerov method (80) applied to equation (79) yields:

$$\underbrace{(M+\frac{\Delta t^2}{12}K)}_{=A}u^{n+1} = \underbrace{\left(2M-\frac{5\Delta t^2}{6}K\right)u^{n}}_{=v_1} - \left(M+\frac{\Delta t^2}{12}K)\right)u^{n-1}. \tag{81}$$

This defines a system of equations which are solved by first rearranging the terms in (81):

$$A\underbrace{(u_{n+1}+u_{n-1})}_{=v_2} = v_1 \tag{82}$$

and then solving (82) for v_2 and extracting u_{n+1} as $u_{n+1} = -u_{n-1} + v_2$.

Solving for v_2 in (82) is the most costly operation and must therefore be optimized as much as possible. The matrix A is symmetric and positive definite, since it is the sum of two positive definite matrices. It can therefore be Cholesky factorized into an upper and a lower matrix G, i.e. $A = GG^T$ [3]. The Cholesky factorization of A is performed once as an initial step in the computations. Forward and backward substitutions give the vector v_2. Because of the sparse structure of the matrices only the seven non-zero diagonals are stored in a matrix with seven columns and routines to do sparse matrix-vector multiplication were implemented.

It is necessary to find some kind of estimate on the restrictions on the time-step Δt in relation to the space-step Δx to achieve stability. Equation (81) on one-step form is:

$$\begin{aligned}&\begin{pmatrix} M+\frac{\Delta t^2}{12}K & 0 \\ 0 & I \end{pmatrix}\begin{pmatrix} u^{n+1} \\ u^{n} \end{pmatrix} = \\ &= \begin{pmatrix} 2M-\frac{5\Delta t^2}{6}K & -(M+\frac{\Delta t^2}{12}K) \\ I & 0 \end{pmatrix}\begin{pmatrix} u^{n} \\ u^{n-1} \end{pmatrix}\end{aligned} \tag{83}$$

With

$$B = \begin{pmatrix} (M + \frac{\Delta t^2}{12}K)^{-1})(2M - \frac{5\Delta t^2}{6}K) & -I \\ I & 0 \end{pmatrix}, \quad \begin{pmatrix} u^{n+1} \\ u^n \end{pmatrix} = v^{n+1} \qquad (84)$$

we can rewrite equation (83) and get $v^{n+1} = Bv^n$.

The eigenvalues of the matrix B will set the stability criterion for the time integration. Let $\begin{pmatrix} a & b \end{pmatrix}^T$ be a vector composed of the vectors a and b. We have that

$$B \begin{pmatrix} a \\ b \end{pmatrix} = \lambda \begin{pmatrix} a \\ b \end{pmatrix} \qquad (85)$$

From equation (84) and (85) we get $a = \lambda b$ and

$$\underbrace{\left(2 - \lambda - \frac{1}{\lambda}\right)}_{=\mu} a = \left(\frac{11}{12}\Delta t^2 M^{-1} K\right) a.$$

Note that μ is the eigenvalue to $\frac{11}{12}\Delta t^2 M^{-1}K$ and that the eigenvalues are the same as in the analysis of the second order time integrator and we can use the result that the largest eigenvalue is obtained for $0 \leq \mu \leq 4$.

An estimate of the largest eigenvalue of $M^{-1}K = \frac{1}{\Delta x^2}\tilde{M}^{-1}\tilde{K}$ is needed. In the "tilde"-matrices, Δx has been factored out. Numerical tests for finer and finer gridresolutions has been performed and indicate that the largest eigenvalue is close to 18.9026. We combine this result with the upper limit $\mu \leq 4$ to obtain

$$\frac{\Delta t}{\Delta x} < \sqrt{\frac{48}{11 * 18.9026}} \approx 0.48.$$

This is a stability criterion for the fourth order FEM approximation with the fourth order Numerov time-integrator with periodic boundary conditions and give an estimate of the time-step one can use in the non-periodic case.

6 Numerical experiments and results

6.1 General remarks and definitions

Throughout this section the following notation for the different programs will be used:

- YEE for the second order Yee finite difference program
- SBP2 for the second order summation by parts program
- SBP4 for the fourth order summation by parts program
- SBP6 for the sixth order summation by parts program
- FEM2 for the second order finite element program
- FEM4 for the fourth order finite element program

Note that for SBP2-SBP6, the fourth order accurate Runge-Kutta method was used for the time integration. The expression PPW [8], should be read as *Points Per Wavelength* and is a way of expressing how many points that is used for each wavelength. Unless stated otherwise the experiments have been performed with five wavelengths inside the computational domain [0,1].

In all experiments a scaled norm of the error has been used to evaluate the results from the different methods,

$$||u - v|| = \sqrt{(u - v)^T A(u - v)}$$

where u is the numerical solution and v is the exact solution projected onto the grid. In the YEE method the matrix A is the identity matrix scaled with Δx. In the summation by parts methods, A is the matrix P (see section 4.2). For the FEM methods the matrix A is replaced by the matrix M (see section 5.3 and 5.4). From here on the above stated measures of the error in different norms will be referred to as "the norm of the error".

For each method, there is a bound on the allowed time-step for stability, see table 1 . The relation between the bound on the time-step and the time-step actually used in the calculations is usually referred to as the CFL-condition (Courant-Friedrichs-Lewy). E.g., CFL=0.9 means that the time-step used is 90% of the largest time-step allowed for stability. For the YEE method the stability limit is determined by the limit in the analysis of the periodic problem (see section 3.2). For the finite element methods, estimates of the stability limits are derived in sections 5.3 and 5.4. The stability limits for the SBP methods were determined using the stability limitations of the fourth order Runge-Kutta time-integrator and an estimate of the largest eigenvalue of $P^{-1}Q$ (see eq. (43)).

Table 1. The allowed timestep due to stability limits.

Method	$(\Delta t/\Delta x) <$
YEE	1
SBP2	1.97
SBP4	1.11
SBP6	0.72
FEM2	0.58
FEM4	0.48

6.2 Equivalence between YEE and SBP2

The Yee scheme and the second order accurate SBP scheme are equivalent on the semi-discrete level, when using twice as many points in the SBP scheme as in the Yee scheme. In fig. 9 the Yee-molecule is used on the left side and the second order SBP-operator (D_0) is used on the right side to calculate the next H-value. A test was performed using a small Δt to eliminate the temporal error and the calculation was terminated at $T = 1$. The result is presented in table 2.

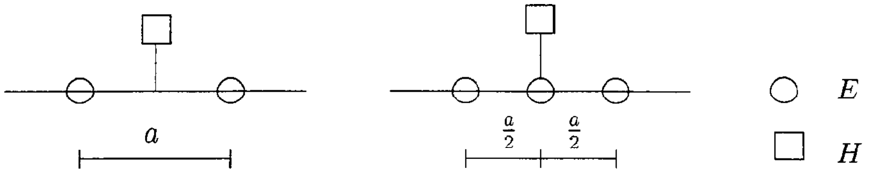


Figure 9. Left: Yee with spatial step-length a. Right: SBP2 with spatial step-length $\frac{a}{2}$.

Table 2. Equivalence between YEE and SBP2 in the spatial dimension.

# Points	Max Norm YEE	# Points	Max Norm SBP2
100	8.5e-2	200	8.8e-2
200	2.1e-2	400	2.2e-2
400	5.2e-3	800	5.5e-3

From table 2 we see that the spatial discretizations seems to be equivalent.

6.3 Order of accuracy

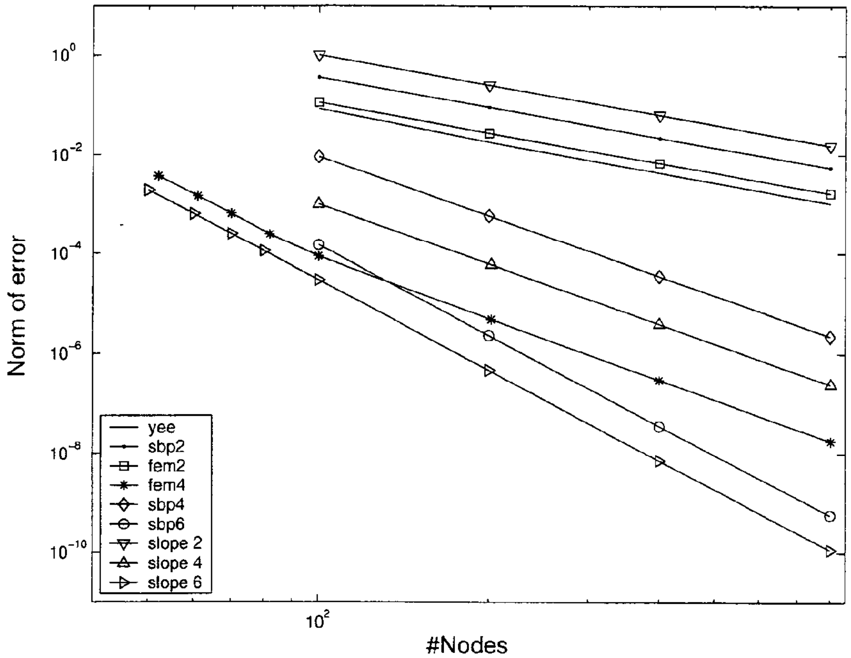
To check the order of accuracy in time we keep Δx constant and let $u^n = u_{\text{exact}} + C_1 \Delta x^p + C_2 \Delta t^q$. We use the formula

$$\frac{u^{\Delta x, \Delta t} - u^{\Delta x, \frac{\Delta t}{2}}}{u^{\Delta x, \frac{\Delta t}{2}} - u^{\Delta x, \frac{\Delta t}{4}}} = 2^q,$$

where the spatial parts have canceled out and q is the order of accuracy in time. For the second order methods YEE and FEM2, q is near 2 and for SBP2, SBP4, SBP6 and FEM4 q is 4.

The order of accuracy in space was investigated by examining how the norm of the error behaves as the resolution in the domain is increased. A series of tests was performed and the norm of the error was plotted as a

Figure 10. Order of accuracy is determined. CFL is 0.9 for all methods except FEM4 and SBP6 which have CFL=0.1



function of the number of grid-points in the computational domain. Figure 10 show that all methods have the expected order of accuracy. The CFL-number was 0.9 in all calculations except for the SBP6 and the FEM4 cases. In the SBP6 case, a low CFL-number is used to remove the temporal errors introduced by the fourth order Runge-Kutta method. The FEM4 method show a super convergent behavior for low grid resolutions (50-100 grid-points, sixth order in space). For really small Δx, the fourth order accuracy in space was verified.

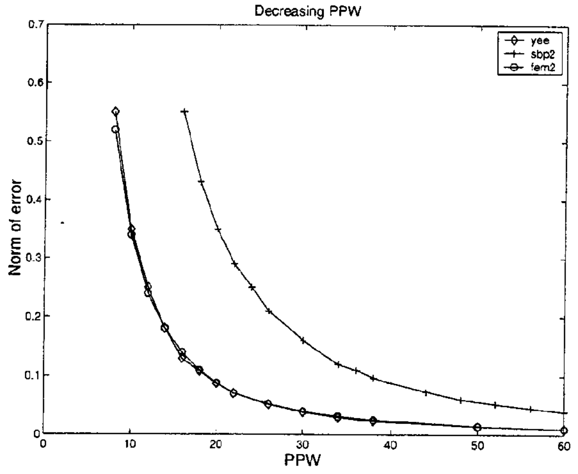
Note also the increase in accuracy with the higher order methods compared to the lower order methods. It takes at least four times as many grid-points with a second order method to achieve the same error as with a fourth order method.

6.4 Brakedown

When propagating electromagnetic waves over long distances, it is important to know how many grid-points must at least be used without loosing all information in the wave.

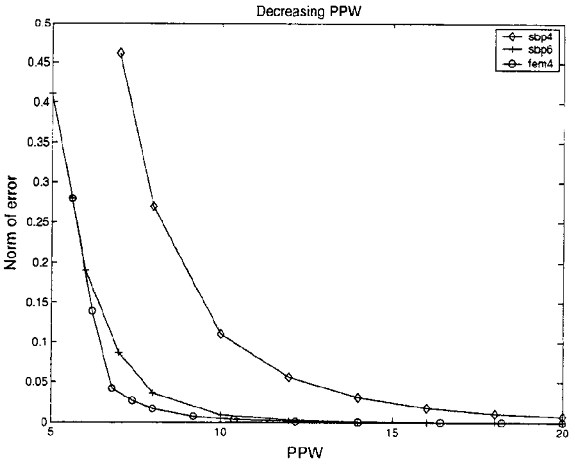
To answer this question, the "Brakedown" tests below were performed. The idea behind these tests was to let an electromagnetic wave propagate a certain distance (calculations were performed to $T = 1$) and use fewer and fewer grid-points in the computational domain until the method brakes down and the error is too large for the result to be useful. The norm of the error is plotted for each resolution. Note that in figure 11 the YEE method is plotted with the number of E-nodes used for each wavelength. If one would plot the YEE method in the same figure but count both E and H-

Figure 11. Brakedown of the second order methods.



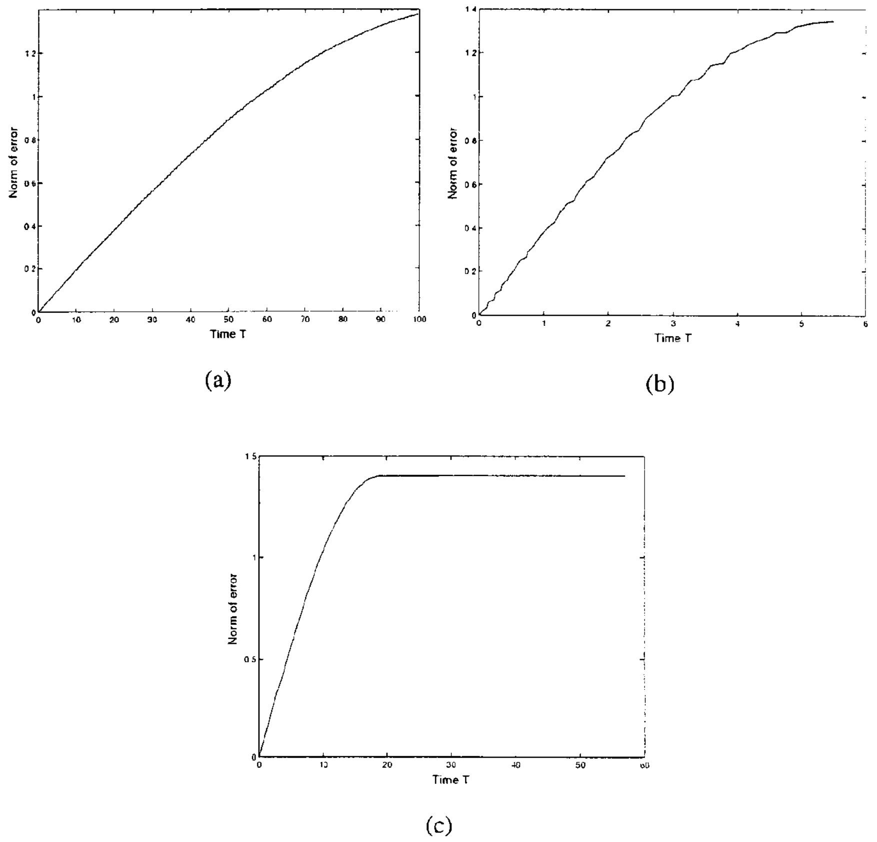
values for each wavelength, the YEE and SBP2 methods are very much alike. It is clear though, that the FEM2 method needs fewer PPW than the SBP2 method. In fig. 12, SBP6 and FEM4 show very low errors down to 10 PPW but SBP4 needs almost twice as many PPW to achieve the same accuracy. That the SBP6 is better than SBP4 is no surprise. The reason that the FEM4 method can compete with, or even outperform, the SBP6 method is the super-convergence for few grid-points mentioned in section 6.3 above.

Figure 12. Brakedown of the fourth order methods and the sixth order summation by parts method.



6.5 Long Time-integration

To investigate how the norm of the error develops for large times, tests have been performed with different grid-resolution to $T = 100$. Calculations with grid-resolutions 20 and 80 PPW are displayed for all methods. We measure the norm of the error and save only the maximum value as it is increased in time. Figure 13 show that 20 PPW is not sufficient for time integration over such long time periods as $T = 100$ when using second order methods. YEE reaches it's maximum error at approximately $T = 100$ but SBP2 reaches it's maximum at approximately $T = 6$. For FEM2, the time is approximately $T = 20$. Figure 14 show that 80 PPW does not suffice to keep the norm of the error down. With 80 PPW, YEE can almost keep the norm of the error below 10% at $T = 100$ but FEM2 and SBP2 is not even close to that level. The norm of the error increases above 10% approximately at $T = 20$ with the FEM2 method.



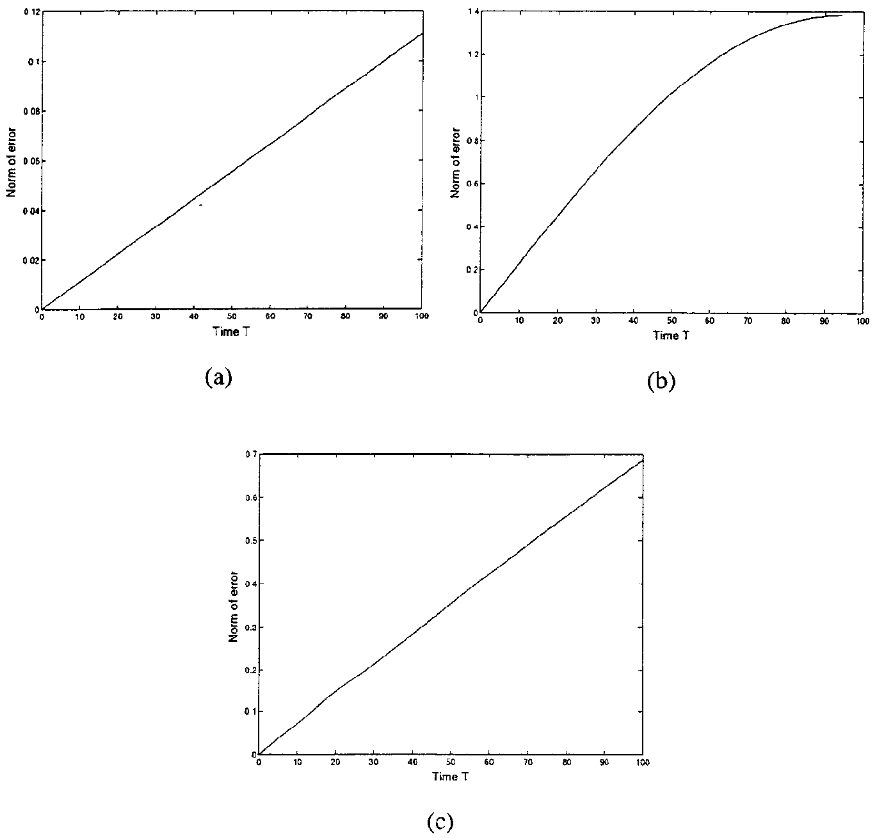
(a)

(b)

(c)

Figure 13. (a) YEE 20 PPW (E-points), (b) SBP2 20 PPW, (c) FEM2 20 PPW

Figure 14. (a) YEE 80 PPW, (b) SBP2 80 PPW, (c) FEM2 80 PPW

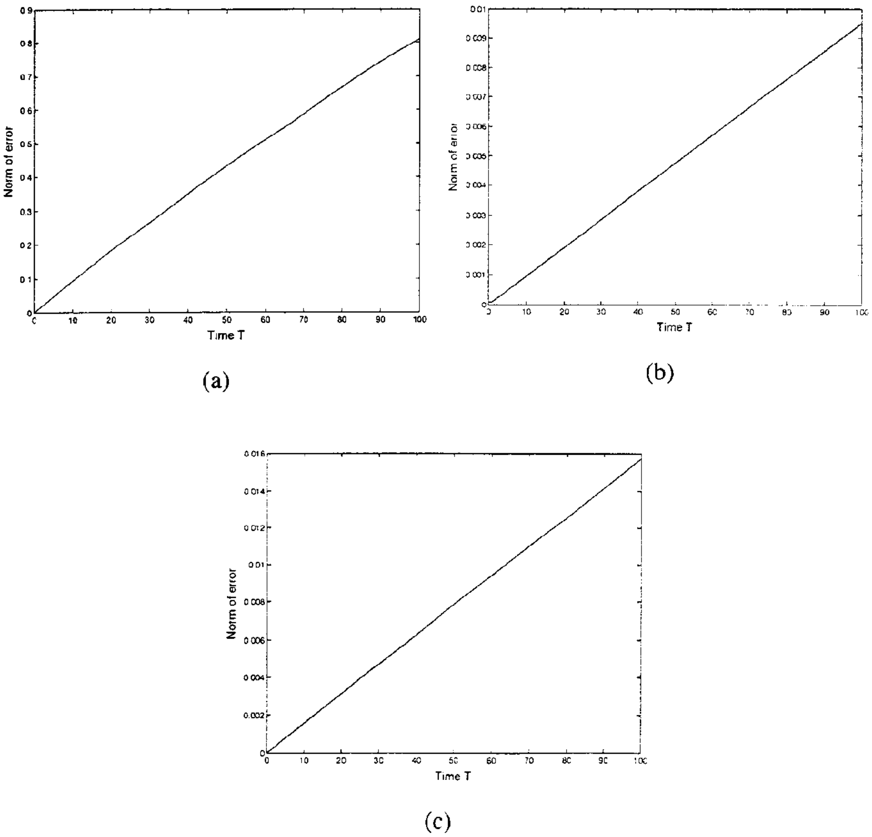


(a)

(b)

(c)

Figure 15. (a) SBP4 20 PPW, (b) FEM4 20 PPW, (c) SBP6 20 PPW



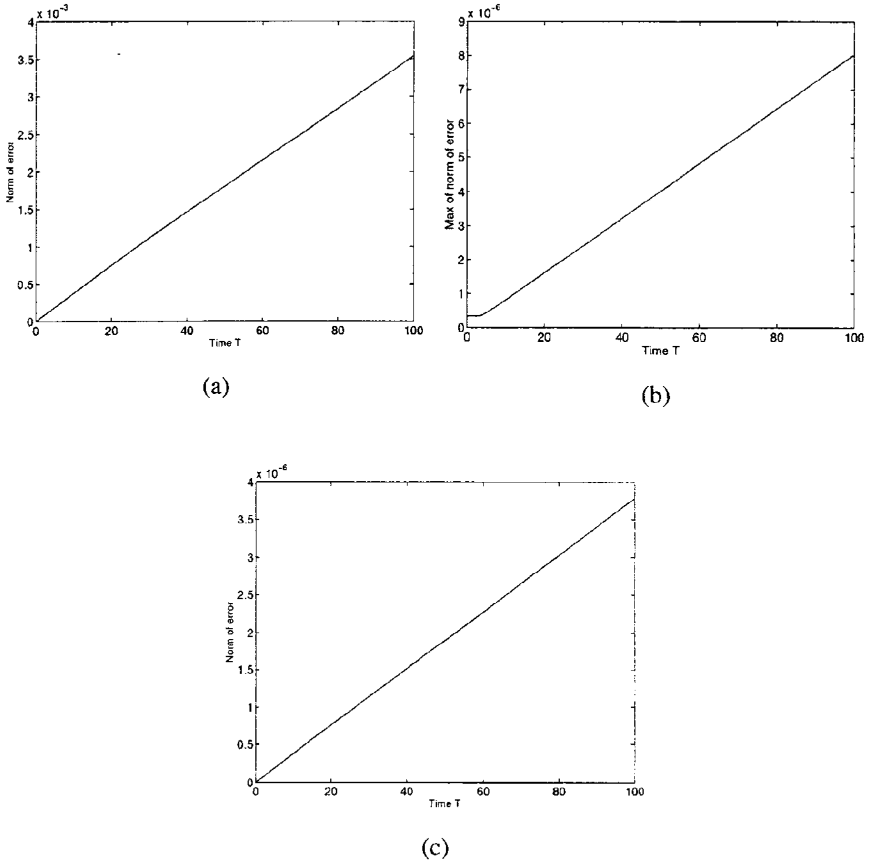
(a)

(b)

(c)

When using 20 PPW FEM4 and SBP6 can keep the norm of the error fairly low (<1%) at $T = 100$ but SBP4 can not do that. The norm of the error is more than 80% at $T = 100$, see fig. 15.

Figure 16. (a) SBP4 80 PPW, (b) FEM4 80 PPW, (c) SBP6 80 PPW



(a)

(b)

(c)

With 80 PPW, SBP4 manage to keep the norm of the error below $\approx 10^{-3}$ at $T = 100$ but the fourth order FEM4 and the sixth order SBP6 keeps the norm of the error below $\approx 10^{-6}$. Note that, using 80 PPW with SBP4 keeps the norm of the error below what is achieved with 20 PPW with FEM4 and SBP6, see fig. 15 and 16.

6.6 Error under a limit

In previous sections, the accuracy of the methods have been studied but we have not yet considered the efficiency. An efficient numerical method must produce an accurate result in a reasonable time. The experiments below, ("error under a limit"), show for each method, how many PPW that is needed to keep the norm of the error below a predefined limit. We perform calculations with each method and increase the number of grid-points until we can keep the norm of the error below the limit during the whole time integration. Note that the results from the YEE method is here presented with PPW meaning the number of E-nodes that has been used per wavelength. In the figures 17 and 18, T_c denotes the CPU-time needed to achieve the desired accuracy of 1 or 5%. Note that for low grid-resolutions the FEM4 method is in the super-convergent area. For all methods CFL=0.9 has been used except for SBP6 were CFL=0.1 was used due to the accuracy limitations of the fourth-order time-integrator.

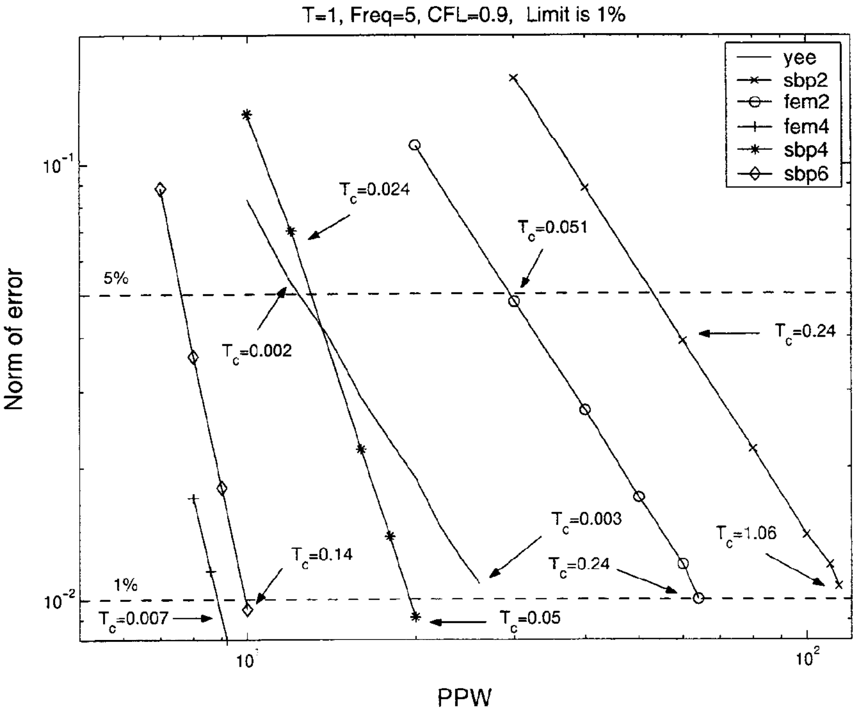
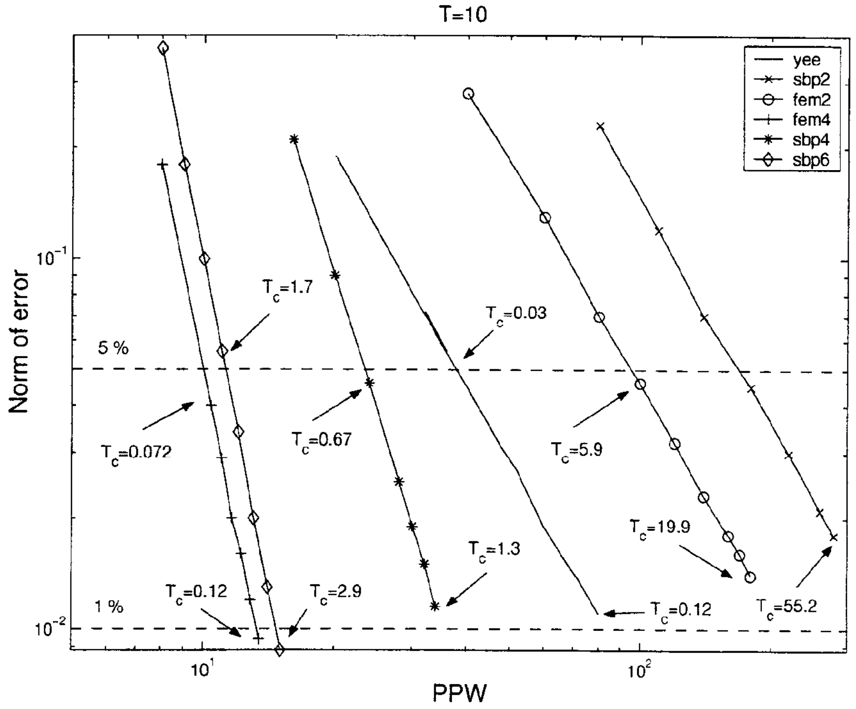


Figure 17. PPW needed to keep the norm of the error under a specified limit during the time integration to T=1.

When performing time-integration to $T = 1$ we see in fig. 17 that the YEE method uses less PPW than SBP2 and FEM2 and the CPU-time to achieve the 1% limit is much smaller than for the other two second order methods. The efficiency of YEE is due to the staggered formulation of the method. It uses only half the number of unknowns compared with the SBP2 method. The leap-frog time integration is also more efficient than the Runge-Kutta method used in the SBP2. The FEM2 method is quite efficient, not close to YEE, but better than SBP2.

For the fourth order methods we see in fig. 17 that SBP4 is not as fast as FEM4 (which is in the super-convergent area) or the SBP6 but we see that it outperforms both SBP2 and FEM2. The desired accuracy is achieved much faster than for the two second order methods. The YEE method though is faster but uses a little more grid-points. FEM4 and SBP6 uses so few PPW that it is difficult to compare them at the time $T = 1$.

Figure 18. PPW needed to keep the norm of the error under a specified limit during the time integration to T=10.



When performing the calculations to $T = 10$ (see fig. 18), we see that the YEE method is far better than the other two second order methods, SBP2 and FEM2. The simplicity of YEE makes it fast and the CPU-time does not increase as much as for the other two methods. For YEE the CPU-time increases by a factor 40, for SBP2 by 55 and for FEM2 by over 80 from the case $T = 1$ to the case $T = 10$.

For the fourth and sixth order methods we see in fig. 18 that FEM4 is again very efficient. It uses fewer grid-points than SBP6 and is faster. But the SBP6 can not perform efficiently since we use a Runge-Kutta of order four for the time-integration.

To make another comparison of the efficiency of the fourth order methods, an additional test has been made. By requiring an error far below the previous used limit of 1%, at $T = 1$, we force both methods to use more gridpoints and FEM4 will then be out of the super-convergent area. The limit chosen is a little greater than 10^{-7}.

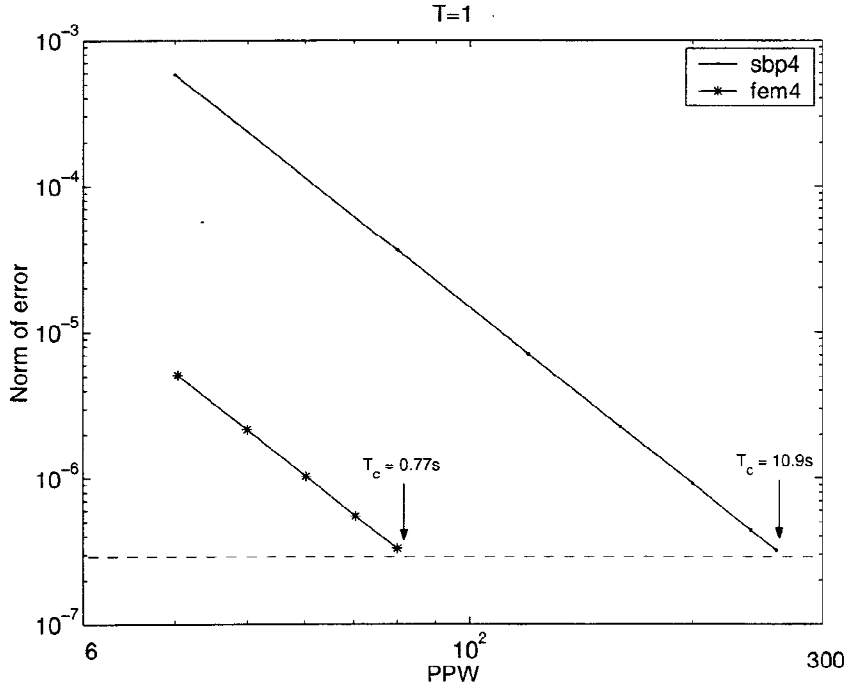


Figure 19. PPW needed to keep the norm of the error under a specified limit during the time integration to T=1. Fourth order methods

In fig. 19 calculations have been performed to $T = 1$ as in the previous test displayed in fig. 17. Note that the CPU-times T_c in fig. 19 should not be compared with the CPU-times in fig. 17 directly since different computers were used in the two tests. We see in fig. 19 that FEM4 needs only 80 PPW while SBP4 needs almost 300 PPW to achieve the desired accuracy at $T = 1$. The error in the finite element method is approximately 100 times lower than the error with the summation by parts method for the same grid-resolution. This means that more grid-points are needed and efficiency is lost. The CPU-time is more than 14 times as large for the SBP4 compared to the FEM4 method.

As a comparison, the second order YEE method was used to achieve the same low limit of the norm of the error. It turns out that YEE needs more than 1600 PPW and a CPU-time over 147s to achieve this low limit.

7 Extension to higher dimensions

As mentioned in the introduction, a 1D analysis provides most of the answers required for choosing a computational method, except for the question about scaling the efficiency from 1D to 3D. We do the extension to a higher dimension by estimating the computational cost in arithmetic operations needed to update the unknown variables in one grid-point in a 1D and a 3D problem. By dividing the number of arithmetic operations in 3D with the number of arithmetic operations in 1D we get a factor which indicates how the efficiency scales from 1D to 3D, see table 3.

Since the Yee-scheme is staggered in space and time the approach has been to calculate the cost to update the six unknowns, three of the electric field and three of the magnetic field, that uniquely defines a Yee-cell in 3D. In 1D, only two unknowns are unique in each Yee-cell. In the SBP finite difference methods there are two unknowns in each grid-point in 1D and six in 3D.

The FEM have one electric field component in each grid-point in 1D and three components in 3D. For the Yee-scheme and SBP we can calculate the computational cost exactly but for FEM we will make a rough estimate of the cost of solving the system of equations arising from the implicit formulation.

We assume a uniform grid of tetrahedras with N_1, N_2 and N_3 grid-points in the x-, y- and z-direction respectively. Let $M_1 = 3N_1$ denote the unknowns in the x-direction, $M_2 = 3N_2$ denote the unknowns in the y-direction and so on. The bandwidth can be reduced from $M_1M_2M_3$ to M_1M_2 by properly numbering the unknowns. The banded system of equations that arise is LU-factorized only once at a cost of $\mathcal{O}(2npq)$ arithmetic operations, where n is the number of unknowns and p and q is the lower and upper bandwidth respectively (see [3]). In this case the number of unknowns is $M_1M_2M_3$ and the bandwidth is approximately $\frac{M_1M_2}{2}$ (both upper and lower) which results in $\mathcal{O}(M_1M_2M_3M_1^2M_2^2)$ arithmetic operations to LU-factorize the system. The cost of solving the LU-factorized system in each time-step is approximately $2np$ arithmetic operations for the forward substitution and $2nq$ for the backward substitution (see [3]). This results in $\mathcal{O}(\underbrace{M_1M_2M_3}_{Unknowns}\ \underbrace{M_1M_2}_{Bandwidth}\ )$ arithmetic operations in our case. The bandwidth is larger for third order polynomials, but differs only by multiplication of a constant, which we will denote by α. In table (3) we have divided the cost of solving the system with the number of unknowns to estimate the cost of updating the unknowns in one grid-point. We also suppose that we have the same number of unknowns in each direction in the grid, i.e. $M_1 = M_2 = M_3 = M$.

In 1D, the FEM methods give rise to narrow banded systems of equations, (seven non-zero diagonals in the FEM4 case), which explain the low number of arithmetic operations needed to update one grid-point.

Table 3. The number of arithmetic operations needed to update one grid-point indicating the efficiency when scaling from 1D to 3D

Method	1D	3D	Factor
YEE	6	36	6
SBP2	38	258	6.8
SBP4	54	402	7.4
SBP6	70	546	7.8
FEM2	18	$\mathcal{O}(M^2)$	$\mathcal{O}(M^2)$
FEM4	114	$\mathcal{O}(\alpha M^2)$	$\mathcal{O}(\alpha M^2)$

For the FEM method a direct LU-solver was considered in the 3D case. However, if a good preconditioner can be found, an iterative solver could improve the efficiency of 3D FEM in table 3.

8 Conclusions

For second order methods, SBP2 cannot compete against the other two second order methods neither if the criteria is accuracy nor efficiency. FEM2 and YEE are equally accurate but YEE would be the final choice because of better efficiency.

For fourth order methods FEM4 is almost a factor 100 more accurate than SBP4. When FEM4 is in the super-convergent area it can compete with SBP6 in accuracy. The FEM4 method is also more efficient than SBP4 outside the super-convergent area.

Higher order methods produce more accurate results than second order methods when using the same number of PPW. Higher order methods are also better at transporting waves using low PPW numbers for long time-integrations. YEE can compete in efficiency with higher order methods for short time-integrations and high accuracy limits. With a low accuracy limit, the higher order methods outperform YEE.

In 3D the FEM is less efficient than the other methods due to the implicit formulation which results in a large system of equations that must be solved using direct methods. With an efficient iterative solver, FEM may be able to compete with the other methods because of its supreme accuracy.

Appendix A

Kronecker product

Definition 1 *Let $\boldsymbol{A}$ be a $p \times q$ matrix and let $\boldsymbol{B}$ be an $m \times n$ matrix, then*

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{0,0}\boldsymbol{B} & \cdots & a_{0,q-1}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ a_{p-1,0}\boldsymbol{B} & \cdots & a_{p-1,0}\boldsymbol{B} \end{pmatrix} \tag{86}$$

The $p \times q$ block matrix $\boldsymbol{A} \otimes \boldsymbol{B}$ is called a Kronecker product.

There are a number of rules for Kronecker products, see [11], we will present some of them. Let $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$ be matrices of arbitrary sizes, such that the specified operations are defined.

$$\begin{gathered} (\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = (\boldsymbol{AC}) \otimes (\boldsymbol{BD}) \\ (\boldsymbol{A} + \boldsymbol{B}) \otimes \boldsymbol{C} = \boldsymbol{A} \otimes \boldsymbol{C} + \boldsymbol{B} \otimes \boldsymbol{C} \\ (\boldsymbol{A} \otimes \boldsymbol{B})^T = \boldsymbol{A}^T \otimes \boldsymbol{B}^T \\ (\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1} \\ \boldsymbol{A} > 0, \quad \boldsymbol{B} > 0 \Rightarrow (\boldsymbol{A} \otimes \boldsymbol{B}) > 0 \end{gathered} \tag{87}$$

References

[1] M.H. Carpenter, D. Gottlieb, and S. Abarbanel. Time-stable boundary conditions for finite-difference schemes solving hyperbolic systems:methodology and applications to high-order compact schemes. *Journal of Computational Physics*, pages 111:220–236, 1994.

[2] David K. Cheng. *Field and Wave Electromagnetics.* Addison-Wesley, second edition, 1992.

[3] G. Golub and C. Van Loan. *Matrix Computations.* The Johns Hopkins University Press Ltd., third edition, 1996.

[4] B. Gustafsson, H.-O. Kreiss, and J. Oliger. *Time Dependent Problems and Difference Methods.* John Wiley & Sons, 1995.

[5] J.M Jin. *The Finite Element Method in Electromagnetics.* John Wiley & Sons, 1993.

[6] C. Johnson. *Numerical Solution of Partial Differential Equations by the Finite Element Method.* Studentlitteratur, Lund, 1987.

[7] H.-O. Kreiss and G Scherer. Finite element and finite difference methods for hyperbolic partial differential equations. *Mathematical Aspects of Finite Elements in Partial Differential Equations*, 1974.

[8] D.P. Lockard, K.S. Brentner, and H.L. Atkins. High accuracy algorithms for computational aeroacoustics. *AIAA J.*, 1994.

[9] R.D. Skeel, G. Zhang, and T. Schlik. A family of symplectic integrators: Stability, accuracy, and dynamics applications. *SIAM Journal of Comp.*, vol. 18, 1997.

[10] A. Taflove. *Computational Electromagnetics, The Finite-Difference Time-Domain Method.* Artech House, Boston, 1995.

[11] C. Van Loan. Computational frameworks for the fast fourier transform. *SIAM*, 1992.

[12] K.S. Yee. Numerical solution of initial boundary value problems involving maxwell's equations in isotropic media. *IEEE Transactions on Antennas and Propagation*, vol. 14(3):302–307, 1966.

Issuing organisation FOI – Swedish Defence Research Agency Division of Aeronautics, FFA SE-172 90 STOCKHOLM	Report number, ISRN FOI-R–0119–SE	Report type Scientific Report
	Month year May 2001	Project number E840243
	Customers code 3. Aeronautical Research	
	Research area code 6. Electric Warfare	
	Sub area code 62. Stealth Technology	
Author(s) Jonas Persson and Jan Nordström	Project manager Jan Nordström	
	Approved by Torsten Berglind Head, Computational Aerodynamics Department	
	Scientifically and technically responsible Jan Nordström	

Report title

Discrete Approximations of Electromagnetic Problems

Abstract

In this report second and higher order methods (the Yee-method, Summation By Parts methods and Finite Element Methods) for transportation of electromagnetic waves are compared. Tests of accuracy, long time-integration and efficiency are performed. We show that the higher order methods in almost every case outperform the second order methods.

Keywords

Finite difference, Yee, Summation By Parts, Finite Element Method, accuracy, efficiency

Further bibliographic information

ISSN ISSN 1650-1942	Pages 59	Language English
	Price Acc. to price list	
	Security classification Unclassified	

Utgivare
Totalförsvarets Forskningsinstitut – FOI
Avdelningen för Flygteknik, FFA
SE-172 90 STOCKHOLM

Rapportnummer, ISRN FOI-R-0119-SE	Klassificering Vetenskaplig rapport
Månad år Maj 2001	Projektnummer E840243

Verksamhetsgren
3. Flygteknisk forskning

Forskningsområde
6. Telekrig

Delområde
62. Signaturanpassning

Författare
Jonas Persson och Jan Nordström

Projektledare
Jan Nordström

Godkänd av
Torsten Berglind
Chef, Institutionen för Beräkningsaerodynamik

Tekniskt och/eller vetenskapligt ansvarig
Jan Nordström

Rapporttitel
Diskreta approximationer av elektromagnetiska problem

Sammanfattning

I denna rapport jämförs andra och högre ordningens metoder (Yee-metoden, Summation By Parts metoder och Finita Element Metoder) för transport av electromagnetiska vågor. Tester av noggrannhet, lång tids-integration och effektivitet hos de olika metoderna har utförts. Vi visar att i nästan alla fall är högre ordningens metoder bättre än andra ordningens metoder.

Nyckelord
Finita differenser, Yee, Summation By Parts, Finita Element Metoden, noggrannhet, effektivitet

Övriga bibliografiska uppgifter
Examensarbete

ISSN ISSN 1650-1942	Antal sidor 59	Språk Engelska

Distribution enligt missiv

Pris
Enligt prislista

Sekretess
Öppen